

## ABSTRACT

Title of dissertation:     **STATISTICAL KNOWLEDGE AND LEARNING  
IN PHONOLOGY**

Ewan Michael Dunbar, Doctor of Philosophy, 2013

Dissertation directed by:   **Professor William Idsardi  
Department of Linguistics**

This dissertation deals with the theory of the phonetic component of grammar in a formal probabilistic inference framework: (1) it has been recognized since the beginning of generative phonology that some language-specific phonetic implementation is actually context-dependent, and thus it can be said that there are gradient “phonetic processes” in grammar in addition to categorical “phonological processes.” However, no explicit theory has been developed to characterize these processes. Meanwhile, (2) it is understood that language acquisition and perception are both really informed guesswork: the result of both types of inference can be reasonably thought to be a less-than-perfect commitment, with multiple candidate grammars or parses considered and each associated with some degree of credence. Previous research has used probability theory to formalize these inferences in implemented computational models, especially in phonetics and phonology. In this role, computational models serve to demonstrate the existence of working learning/perception/parsing systems assuming a faithful implementation of one particular theory of human language, and are not intended to adjudicate whether that theory is correct. The current dissertation (1) develops a theory of the phonetic component of grammar and how

it relates to the greater phonological system and (2) uses a formal Bayesian treatment of learning to evaluate this theory of the phonological architecture and for making predictions about how the resulting grammars will be organized. The coarse description of the consequence for linguistic theory is that the processes we think of as “allophonic” are actually language-specific, gradient phonetic processes, assigned to the phonetic component of grammar; strict allophones have no representation in the output of the categorical phonological grammar.

STATISTICAL KNOWLEDGE AND LEARNING  
IN PHONOLOGY

by

Ewan Michael Dunbar

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2013

Advisory Committee:

Professor William Idsardi, Chair, Advisor

Professor Hal Daumé III

Professor Naomi Feldman, Co-Advisor

Professor Norbert Hornstein

Professor Jeff Lidz

Professor Rochelle Newman, Dean's Representative

© Copyright by  
Ewan Michael Dunbar  
2013

## Dedication

For Chanelle.

## Acknowledgments

Tell me that it's a wonder, so that I may sleep when all I see in the night is this place. — Ambassador Delenn, “A Voice in the Wilderness, Part 2”

The last five years of my life have been important. I wish I had the space to write the human story of this dissertation. If I do write that story some day, I hope it will carry a lesson I could have benefitted from knowing earlier: the “exercise” model of intellectual work and development is dead wrong. Those who tell you to “push yourself harder” to “get the job done” are sending you the wrong message, unless they are simply trying to make you quit. If this message works, it works largely by accident. This is not to say that you should not find ways to get the job done effectively. But over the last five years, I have flourished whenever I have removed barriers to the contented flow of work, not pushed against them, and the most effective way of doing that has been to remember that I am doing this because it is a part of my life. That story ends with a line something like, “You are more than your work, and your work will never be more than you—do not let unhappy and confused people tell you otherwise.” I may not be the person to write it. Still, some of the people I owe the greatest debt to are the ones who have helped me to understand this.

Intellectually, this work began to coagulate back in 2008 around an insight of Brian Dillon's, to whose creativity and vision and ability to intuitively grasp technical problems I owe a great debt. My advisor, Bill Idsardi, was key in taking an insight, turning it into a project, and then guiding the project into coherent ideas. But “coherent ideas” is understating it with Bill. Having a clear vision of the whole field—theory from soup to nuts, integrated with a real theory of inference—is more like it, and is rare enough. The persistence and the audacity with which you reminded me that it can actually be executed is of a higher order still. This has been one of the major forces motivating me over the last five years. And, although the official style sheet indicates that committee member names after the chair must be listed in alphabetical order, Naomi Feldman has been co-advisor since she arrived, to the extent that my unpredictable ventures from the cave to receive advice have supported it. Thanks for always seeing my quest for accuracy and raising me common sense—and for your reminders to stop working. Jeff Lidz and Hal Daumé have, in addition to serving as committee members, always reminded me that if you cannot explain it simply, you have not understood it properly. And, as is often the case, I would like to thank Norbert Hornstein for general inspiration—but, in this case, also for volunteering as an eleventh hour replacement committee member. Thanks also to Amy Weinberg and Rochelle Newman for their roles on the committee.

The broader Linguistics Department/CNL Lab group are irreplaceable, and to all those who built the spark and spirit of this place—and it is worth mentioning that it is clear even without turning back time that Colin Phillips has always been instrumental in driving this institutional vision—no thanks will ever be enough. It has a life of its own now. Alexis Wellwood, in addition to being a wonderful friend, stands out as someone who was evidently drawn here not only because she saw the spark, but also because she is a vessel through which this spirit is eager to flow. You have taught me a lot, and you would be surprised at how much of this thesis is made out of the material we have started

to construct together. To our year—Shevaun Lewis, Brad Larson, Dave Kush, Wing-Yee Chow, Terje Lohndal, Michaël Gagnon, Chris LaTerza—listing the individual contributions to the intellectual and interpersonal tenor of the department and to the ideas and work in here—I think this burden goes mainly in the human story. However, suffice it to say, I am better off for having met you all, not only because you are all very smart. Mr Shepard would of course have to learn to draw both Dave and Brad extra well in the human story. Dave, thank you for co-pontificating on matters of all kinds. Within the department, Howard Lasnik also deserves one special, general thanks: your attention to detail is more than a beneficial skill—it is a method and a vision and a way of thinking that has profoundly influenced my work and my teaching. Thanks also to Josh Falk, who is, after a semester of helping me gut it of a lot of cruft and dead ends, a co-author of the Java code used in Chapter 3. Thanks to Jesse Shawl, whose Honors thesis work did not make it to being mentioned in the main text, but was a set of corpus and modelling results that formed the testbed for other versions of the gender experiments. Special thanks to Jorge Tartamudeo, the Basement Lab manager.

Innumerable “outside” people have made useful comments and suggestions relating to this work over the years, ranging from one-offs to extended discussions. I did not take down all their names. Some of the people whose comments, however trivial-seeming some of them might have been to them at the time, were useful, are: Jeff Heinz; Elan Dresher; John Kingston; Andrew Nevins; Jordan Boyd-Graber; John McCarthy; Joe Pater; Adam Albright; Kathleen Currie-Hall; Jason Eisner; and Alan Yu. Special thanks, as always, go to Alana Johns, Derek Denis, and Mark Pollard, for providing Inuktitut data. Thanks to Anja Lindelof at the Greenlandic Broadcasting Corporation for providing Kalaallisut recordings and transcripts.

At the risk of drawing out the articulated web of influence too far, thanks also goes to all those who gave me a good undergraduate and Master’s education at the University of Toronto. In this particular context, the reader is invited to search for the influence of the late Ed Burstynsky, who was the undergraduate advisor when I arrived. He was, I take it, a large part of what built the collegial atmosphere in the Linguistics department that all of us took for granted. Undergrads and grads were all colleagues and friends, and this was a particularly nice oasis to have given the social desert that U of T can be otherwise. His direct influence here is limited to one of his famous lines from LIN100—one that virtually everyone in the department found some reason to repeat—on the topic of phonetic grounding: “We’ve been writing it /p/, but it could have been ‘maple leaf.’”

A tremendous thanks goes to Farah Fossé, who helped the Crittenden house fight an illegal eviction attempt and thereby played a key role in saving me from being homeless during part of the writing of this thesis. You and the offending landlord could also go in the special soul-savers section of these acknowledgments, (see below), but you did not mean to touch me. Nevertheless, you did: Craig by putting up walls of self-serving falsehoods, and Farah by calmly and patiently providing us with the truth. Together, you showed me the power of truth, reason, and independent thought: facts matter. I now have no interest in staying quiet when people are hiding behind being “partly right,” (the most frequent case, unfortunately), and a much clearer head about speaking up. Thanks also to Molly McCullagh and Dave Kush for offering to help.

My close personal relationships with many people have been instrumental in my

work and general survival over the past five years. Some of them have been mentioned already, but many, many others remain. A list, in no particular order, of people who have at least made some small home in my heart within the last five years, if not set up shop there, or come home briefly to resume a long-standing tenancy: Alex Essoe, Dave Kush, Nathan Rolleman, Christina Bjorndahl, Ailis Cournane, Smiljka Tasić, Fernanda Queiros, Sophie Maudslay, Alex “Ace” Carruthers, Emily Oppenheimer, Graham Van Pelt, Nika Mistruzzi, Suzanne Freynik, Amanda Brazeau, Corey Simpson, Kate Borowec, Jack Dylan, and Annie Gagliardi. There are others, of course. My immediate family, Joan Smeaton, Earl Dunbar, and Emily Dunbar, and my extended family, especially my Grandma, Blanche, and late Grandpa, Bill Dunbar, obviously drove who I have become, but also let themselves come to be known to me in new and important ways in the last five years, ways which have shaped me substantially. Special thanks to Callie Wright for keeping me sane during the crunch and to Brock Rough and Darryl McAdams for helping me get my license (all remaining errors in driving are mine).

Finally, to those of you who have done things particular and deliberate which changed me profoundly from the inside during the last five years—at my behest or not—sorry to embarrass, but the deep contributions you made to this dissertation are the underlying content, the rest is details. Normally these special influences and relationships are entered in the acknowledgments with a lot of words, but I can’t do that without turning this into the human story. I separate out the following list of people from those above at great risk in case it looks like the words and hearts of the people I just mentioned have never punched me in the gut. Of course they have, some many times, some for better, some for worse (but ultimately always for better). But, Romy Lassotta, Chiara Frigeni, Laura Broadbent, and Ann Collins, and Sol Lago: I would have fallen apart without you. Well, okay—it might be fairer to say that I did fall apart without you, and you are the reasons Humpty Dumpty is still on the wall and not a thick powdery paste of shell and yolk. Sol had the patience and respect to watch this process with equanimity many, many times, and is perhaps the most sensible person I know. Finally, to Lucile Marteel: I am still waiting on that report.



## Contents

List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 Bayesian inference . . . . .	6
1.2 Phonology . . . . .	13
2 Simplicity	24
2.1 The poverty of the stimulus: what is to be done? . . . . .	24
2.2 Bayesian models of cognition: why should we care? . . . . .	33
2.3 The syntactic acquisition model of Perfors et al. . . . .	42
2.4 Evaluation measures: restrictiveness and simplicity . . . . .	53
2.5 Bayesian Occam's Razor . . . . .	63
2.5.1 Maximum likelihood and restrictiveness . . . . .	63
2.5.2 Model evaluation in statistics . . . . .	67
2.5.3 Bayesian inference and model evaluation . . . . .	74
2.5.4 Conditions for a Bayesian Occam's Razor . . . . .	84
2.6 The Optimal Measure Principle . . . . .	88
2.6.1 Formalizing grammars preliminary: Transparency and structure . . . . .	89
2.6.2 Notation and the structure of a grammar . . . . .	96
2.6.3 Relating grammars to priors in an optimal way . . . . .	109
2.6.4 Example: deriving a symbol-counting evaluation measure . . . . .	111
2.7 Discussion . . . . .	113
2.8 Conclusion . . . . .	116
3 Modelling allophone learning	118
3.1 Categories and transformations . . . . .	118
3.1.1 Empirical review . . . . .	118
3.1.2 Computational and mathematical models . . . . .	124
3.1.3 Phonetic transform hypothesis . . . . .	133
3.2 A computational model: Dillon, Dunbar and Idsardi (2013) . . . . .	144
3.2.1 Mixture of linear models . . . . .	144
3.2.2 Summary of Inuktitut experiments . . . . .	151
3.3 Selecting transform environments . . . . .	158
3.3.1 Mixture of linear models with variable selection . . . . .	158

3.3.2	Experiment: Inuktitut revisited . . . . .	162
3.3.2.1	List of sub-experiments . . . . .	163
3.3.2.2	Results . . . . .	164
3.3.2.3	Discussion . . . . .	168
3.3.3	Experiment: sex and gender differences . . . . .	171
3.4	Proposed model: learning with features . . . . .	175
3.4.1	Background: features, geometries, and the contrastive hierarchy .	175
3.4.2	Background: Bayesian category models with features . . . . .	187
3.4.3	Feature-based phonetic category models: goals for future research	194
3.5	Further issues . . . . .	202
3.6	Summary . . . . .	209
4	Phonetic transforms I: The cognitive architecture . . . . .	211
4.1	The phonetic surface . . . . .	211
4.1.1	Background: Surface representations . . . . .	216
4.1.2	Status of surface representations under a phonetic transform hy-	224
	pothesis . . . . .	
4.1.3	Problematic and unproblematic appeals to surface representations	231
	in phonological theory . . . . .	
4.1.3.1	No problems with AC-representations . . . . .	232
4.1.3.2	Historical changes . . . . .	238
4.1.3.3	Opaque allophony . . . . .	251
4.2	The Lateness of Allophony . . . . .	256
4.2.1	Background: Structure-preservation and the cycle . . . . .	256
4.2.2	Structure-preservation and phonetic transforms . . . . .	263
4.2.3	Issues with phonetic EOD blocks in HOCD theories . . . . .	268
4.3	Summary . . . . .	275
5	Phonetic transforms II: Linguistic phenomena . . . . .	276
5.1	Incomplete neutralization . . . . .	277
5.1.1	Empirical predictions . . . . .	281
5.1.1.1	One category, one process . . . . .	281
5.1.1.2	Two categories, one process . . . . .	286
5.1.1.3	Two categories, one categorical process . . . . .	289
5.1.2	Summary . . . . .	292
5.2	Phonetic process interactions . . . . .	293
5.2.1	Predictions . . . . .	293
5.2.2	Possible counterexamples . . . . .	297
5.3	Statistics in linguistics . . . . .	311
5.4	Main findings . . . . .	315

## List of Tables

2.1	Made-up heights of ten adult Dutch females in centimetres. . . . .	68
3.1	Summary of model evaluations from Dillon, Dunbar & Idsardi 2013 . . .	156
3.2	Complementarity of context distributions of categories estimated by mixture of Gaussians Inuktitut models in Dillon, Dunbar & Idsardi 2013 . . .	157
3.3	Quantitative summary of Experiments 4–11 . . . . .	165
3.4	Difference between the Inuktitut vowel mean conditional on a particular following consonant place and the mean elsewhere, for the four different places of articulation. . . . .	167
3.5	Quantitative summary of Experiments 12–13 . . . . .	172
4.1	The consonant and vowel inventory of Russian . . . . .	238

## List of Figures

2.1	Gaussians of differing variance illustrating restrictiveness in the use of the likelihood-based inference . . . . .	63
2.2	Derivations for the string <i>aaabbb</i> following a context-free grammar and a corresponding categorial grammar . . . . .	98
2.3	Diagram of subpart relations between three CFGs. . . . .	100
2.4	Diagram of subpart relations between four CFGs. . . . .	101
3.1	Mixture of two-dimensional Gaussian distributions . . . . .	127
3.2	Illustrations of conventional mixture of Gaussians models and mixture of linear models as phonetic category systems . . . . .	147
3.3	Second and first formant values for Inuktitut vowel tokens . . . . .	152
3.4	Example fitted models from Experiments 1 and 3 of Dillon, Dunbar & Idsardi 2013 . . . . .	157
3.5	First and second formant values for English corner vowels . . . . .	172
3.6	Example of mixture of linear models fit to English corner vowel data (Experiment 13) . . . . .	173
3.7	Three different ways that the clusters in a three-vowel system might be decomposed if the likelihood function in a latent feature model simply adds Gaussian means . . . . .	192
3.8	One way that the clusters in a three-vowel system might be decomposed if the likelihood function in a latent feature model with feature values $+1$ and $-1$ adds Gaussian means . . . . .	200
3.9	Second and first formant values for Inuktitut vowel tokens . . . . .	202
3.10	Second and first formant values for Kalaallisut vowels. . . . .	205
4.1	Diagram of current versus conventional phonological architectures . . . . .	225
5.1	Duration of voicing and aspiration measurements for German stops in intervocalic, final, and word-initial position (data taken from Jessen 1998 and Port & O'Dell 1986) . . . . .	278

## Chapter 1: Introduction

Mind swept clean like arctic sand. — Bruce Cockburn, “Nanzen Ji”

The purpose of this dissertation is to entrench statistical theory in linguistic theory. I use phonology, and, in particular, a part of linguistic cognition called the “phonetics–phonology interface,” as a model system to illustrate how statistical theory should be used in linguistics.

Statistical theory is about inference, and so is linguistic theory. In Chapter 2, I spell out one particular connection between statistical inference and grammatical inference in general terms, not tied to any particular learning problem. The Bayesian statistical approach to inference corresponds to an “evaluation measure” approach to the language acquisition problem (Chomsky 1965). Given some data and a statement of the set of possible grammars, or more generally the set of possible learnable states, the approach is to state the learner’s relative preferences over those grammars, but to make no commitment to the process by which the change from initial to adult state takes place. The analysis in terms of relative preferences for final states lends itself to general proofs, and I lay out a general version of the Bayesian Occam’s Razor principle, whereby the axioms of Bayesian inference lead inevitably to a preference for grammars/final states which are simpler under weak conditions of hierarchical structure in the set of grammars. This not only derives

the explicit simplicity bias which has been proposed by linguists in evaluation measures for variable-length grammars, but also the bias for analyses that unify lexical items that I argue linguists almost universally and crucially attribute implicitly to the learner. Furthermore, since Bayesian inference can be derived from general axioms of rational inference, I put forward that the predicted biases can and should be used as a tool for empirically evaluating grammatical theories, a tool which is as well or poorly supported as the conjecture that Bayesian statistical theory is sufficient to explain learners' relative preferences for final states.

The rest of the dissertation is an extended example of how to use statistical theory as a theory of learners' preferences and to make it do work in elaborating and reasoning about the empirical consequences of a particular theoretical linguistic proposal. In Chapter 3, I reintroduce the hypothesis that allophony is the result of context-dependent phonetic processes, entertained briefly in the 1980s, but never fully explored. Chapter 3 sets up a theoretical framework for phonetic grammar (but leaves the details of precisely articulating the notion of "allophony" on which the hypothesis turns for Chapter 5). Chapter 3 then presents implemented statistical models for doing inference over phonetic category and process systems, and provides a piece of positive evidence suggesting that the phonetic process treatment of allophonic patterns is better than the usual understanding. The piece of evidence is that the implemented learner does better at finding systems of allophones and from real phonetic data than a minimally different statistical model where allophonic patterns are stated over discrete categories, as in the usual conception. The basic model is re-presented from Dillon, Dunbar, and Idsardi, 2013, and I then present several variants. I present these variants to demonstrate the practice of exploring theoretical proposals by

embedding them in formal statistical inference.

Then I turn to face the consequences of the phonetic theory of allophony. Chapter 4 continues to bracket the details of what exactly constitutes allophony and how exactly it is connected to phonetic processes, but makes explicit the principal architectural consequence if it is true, which may be summed up under the oversimplified slogan, “There is no surface representation.” This is oversimplified because the architecture is still split into two components—a categorical phonological component, and a later phonetic component—and so there is still a “surface” representation, in the sense of “output of the phonological component”—but it is quite far from the “surface.” I distinguish this “abstract category (AC) representation” of the surface from the “surface category (SC) representation” that is usually assumed, and I show what kinds of analyses that make crucial reference to the phonological output are still acceptable, and what kinds of analyses need to be reformulated. One such case, a historical change from East Slavic, (the post-velar fronting), is somewhat complex. Previous analyses not only make crucial reference to SC-representations, but also posit a kind of “instantaneous change” in the language, attributed to learners. Any such radical change from a grammar faithful to the input to one generating an inconsistent pattern is easily shown to be a marked case once we consider the Bayesian formulation of the learning problem: learners quantitatively balance faithfulness to the input (fit, likelihood) and markedness of final states (bias, prior probability). I sketch a new analysis in terms of phonetic enhancement, with phonetic processes as a mechanism, which still attributes faithfulness to the learner.

In the second part of Chapter 4, I show how the shape of the architecture, “phonology feeds phonetics,” can be deduced from the particular representational commitments

of the architecture by deducing a kind of non-recoverability principle. Combined with the attribution of allophony to the phonetic component, this explains the well-known empirical generalization that allophonic rules are “late.” I note, however, that level-ordered architectures, (which are usually the vocabulary under which the “late allophony” generalization is stated) do not interact well with this assumption, under the interpretation that each “level” is divided into an early phonological and a late phonetic component. I show how this could be maintained if non-allophonic information can be still be recovered from the output of the phonetic component.

Finally, in Chapter 5, I turn to focus principally on two empirical patterns, one of which (the existence of incomplete neutralization) is known but was sometimes previously thought to be problematic, and the other of which (simultaneous application of allophonic processes) is a new prediction. I use incomplete neutralization to articulate the association between allophone-like patterning and phonetic processes. The link is not a hard one, but is by way of the learner’s preferences: the absence of a phonetic process implies overlapping distribution, this interacts with the learner’s (simplicity) biases to give detailed predictions about the kinds of corpora that should lead the learner to phonetic transforms, rather than positing separate categories. The use of a formal theory of inference allows “structure-preservation” to be meaningfully reinterpreted with respect to restrictions on outputs (AC-representation) rather than inputs.

In the second part of Chapter 5, I explore the prediction that sets of allophonic processes should show the simultaneous application pattern. The canonical case is the Canadian Raising pattern (counterbleeding; counterfeeding is equally well predicted). This is not problematic the way it is in Optimality Theory, where the relevant grammatical



constraints can only be stated with respect to their outputs. Here, the environments for allophony can and must be stated with respect to the inputs. The theory also deviates from serial approaches since the 1960s, which have uniformly rejected simultaneous application. I argue that the restriction of simultaneous application to allophony solves the undergeneration problem pointed out by Chomsky and Halle, (and predicts that the empirical gap Chomsky and Halle point to in the predictions of simultaneous application is not systematic but accidental). I then explore two closely related problematic cases for this prediction: Polish and Dutch voicing assimilation, which appear to interact with final devoicing in a way inconsistent with simultaneous application. I propose that these involve a late feature deletion operation distinct from the phonetic transforms responsible for the other type of phonetic transforms discussed. This remains outside of the scope of the detailed theory, because the theory does not make use of phonetic features in category representations, but I set the bounds for how it is to interact with phonetic transforms. I leave a worked out theory for future research. I end by summarizing this and other interesting future research projects raised in the course of the dissertation, and I discuss the proper role of statistics in linguistic theory.

For the sake of the linguist reader, I will use the rest of this section to outline the basics of Bayesian statistical inference. For the sake of the computational or mathematical reader, I will use the rest of this section to give the basic empirical outline of the cognitive problem that this dissertation is about (learning phonological systems), and to fill in some basic theory.

## 1.1 Bayesian inference

Although statistics is often understood to be intimately related with quantitative observations, statistical inference is actually more general than that. Under the “subjectivist” view of probability, what is crucially quantitative is an internal, subdoxastic state of the observing agent we call a “degree of belief” or “degree of (rational) credence.” This may be surprising, as many linguists look upon the pushers of statistics with deep suspicion and hostility, and perceive them to be substituting numbers and data for common sense. However, statistics can be pursued as a branch of cognitive science, in which case there is no conflict between statistical inference and standard linguistic theory.

Statistics refers to a family of techniques that make use of probability theory to do inference and make predictions about how things work, based on observations. Statistical inference requires that those observations be delimited very broadly into “events,” but it does not require that the observations be quantitative, or that they ever be directly tabulated as “relative frequencies.” Furthermore, one particular type of statistical inference, Bayesian inference, starts from the assumption that probabilities are in fact in no way grounded in relative frequencies in any finite or hypothetical infinite “population.” The probability of something is, rather, a “degree of belief” attributed to a hypothetical rational agent making predictions about future events, (which will happen to be well-matched by whatever relative frequencies the events turn out to have, under certain general conditions), or adjusting its confidence in particular hypothesized explanations for those events. Applied Bayesian statistics is used for scientific inference, and keeps the rational agent strictly hypothetical—some idealized agent who can tell us what to make of the data in

front of us. Rational models of cognition, on the other hand, make the assumption of a rational agent an empirical conjecture about how the human mind works. The result of taking this approach—whether it is in speech perception, vision, or learning in some particular domain—is inevitably a kind of minimalist/reductionist argument: “Use the principles of Bayesian inference, and you will find that the behavior of humans on Cognitive Problem X follows from the assumption of rationality, and does not require the ad hoc mechanisms previously proposed.”

What does it mean to be rational? Jaynes 2003, roughly following Cox 1946, lays out the following desiderata for a generalization of Aristotelian logic to the case of degrees of belief:

- (1) Degrees of belief are represented by real numbers.
- (2) If some new information  $C$  arises and makes a proposition  $A$  more plausible, but knowing  $A \wedge C$  has no effect on the plausibility of  $B$ , then  $A \wedge B$  must never go down in degree of belief.
- (3) If some new information  $C$  arises and makes a proposition  $A$  more plausible, then it must also make  $\neg A$  less plausible.
- (4) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

From this starting point, Jaynes shows that the following basic axioms hold, where  $C$

stands for “credence”:

$$(5) \quad C[A \wedge B|C] = C[A|B, C] C[B|C]$$

$$= C[B|A, C] C[A|C]$$

$$(6) \quad C[A \vee B|C] = C[A|C] + C[B|C] - C[A \wedge B|C]$$

$$(7) \quad \sum_{i=1}^N C[A_i] = 1, \text{ for mutually exclusive and exhaustive } A_i$$

$$(8) \quad C[A] > 0, \text{ for all } A$$

These in fact constitute one set of axioms for the probability calculus; the alternates all give rise to the same system (although the third is usually extended to countable sets). From now on, I will replace  $C$  with  $\text{Pr}$ , for “probability.” The critical consequence of this is Bayes’ Rule:

$$(9) \quad \text{Pr}[A|B, C] = \frac{\text{Pr}[B|A, C] \text{Pr}[A|C]}{\text{Pr}[B|C]}$$

Or, put in terms that are meaningful for learning, perception, and other kinds of inference:

$$(10) \quad \text{Pr}[\text{Hypothesis}|\text{Data}] = \frac{\text{Pr}[\text{Data}|\text{Hypothesis}] \text{Pr}[\text{Hypothesis}]}{\text{Pr}[\text{Data}]}$$

a.k.a.

$$\text{Posterior probability} = \frac{\text{Likelihood} \cdot \text{Prior probability}}{\int_H \text{Pr}[\text{Data}|h] \text{Pr}[h] \, dh}$$

The power of Bayes’ Rule may not be immediately apparent. The posterior proba-

bility of a particular hypothesis about the correct grammar or model or face or whatever is to be inferred from the input is the degree of belief in the proposition that it was that particular one (the one that would be above labelled  $A$ ). This is the goal of the computation. The prior probability is the organism's inherent bias for one solution over another, stated as probabilities (degree of "prior belief"). The likelihood is a gradient assessment of the degree to which the data is predicted or surprising given the particular hypothesis selected. The denominator is just a normalizing constant that forces the probabilities to sum up to one. (Sometimes we will leave it off and just write  $\propto$ , "proportional to," when it does not matter.) The power is that organisms surely have some inherent relative preferences, (even if these are totally contentless), and are assessing how consistent or surprising some information is all the time. The general principles of reasoning set out by Jaynes as "desiderata" can now be shown to have as a consequence a definite rule about the rational way to update one's beliefs in the face of new information, just on the basis of these two basic functions. The only demand we place on these two functions is that they follow the laws of probability, which is not hard to achieve (though not guaranteed).

The Bayesian decision theory that comes out of this goes like this: pick a loss function,  $L$ . This function tells me, if I know the "correct" answer, how "bad" my solution is. This is different from the prior, which just states my inherent preferences for one solution over another, irrespective of the problem being solved. Then, find a method for using any given data set to construct a solution—call this method  $\theta^*(x)$ —that minimizes the

expected loss, given the available information:<sup>1</sup>

$$(11) \quad \int_{\Theta} \int_X L[\theta, \theta^*(x)] \cdot f(\theta|x) \, dx \, d\theta$$

One such decision rule is the one that simply says “choose the option that has the highest posterior probability.” This option happens to minimize the expected loss if the loss is just wrong = 1, right = 0 (the 0-1 loss). Other loss functions license other strategies (for example, “when the solution is complex, consisting of multiple components, pick the locally best component for each, not the global best”; or “pick the expected value (the average solution) under the posterior”). I will ignore the choice of loss function and continue to just assume the MAP strategy, only because anything else would leave the loss or decision rule as a free parameter—this is what Bayesian cognitive scientists usually do, albeit with no particular justification. In fact, a “full Bayesian” would not in fact “decide on a grammar” at all (in this dissertation, the most frequent case will be “decide on a phonetic category system”). Rather, the rational thing to do, taking into account all possible information, would be, on any given utterance, to infer what was said by averaging the recovered semantic representation over all possible grammars, (weighted, of course, by the posterior distribution given all the utterance previously heard) and indeed averaging over every possible analysis of the utterance within each particular grammar to obtain a single “recovered message”—and weighting each analysis by how likely each

---

<sup>1</sup>The word “expected” just means “averaged using some probabilities as weights, rather than by taking the usual arithmetic mean”; when working with continuous data or hypotheses we use integrals instead of sums and a function  $f$  called the probability density instead of probabilities—this will be touched on briefly in Chapter 2, but can be largely ignored—the point is that we are minimizing the loss averaged over all possible hypotheses and observations.

of the logically possible grammars predicts it to be. In fact, why pick a single “intended message” at all? The “fully Bayesian” result of perception is simply a distribution over possible phonological, syntactic, and semantic representations, as one need only “take an action” when forced to (for example, when one has to generate an utterance). There need not be any competition, however, about what is the “most Bayesian” way of doing things: some part of the brain state may be (isomorphic to) a distribution over states that gets updated following Bayes’ Rule; while some other part of the brain state may not be subject to uncertainty, but may still be updated in a way that is sensitive to a posterior distribution (say, if the “grammar in use” is the best grammar according to the posterior over “candidate grammars”). The question is empirical. For the purposes of this dissertation I hew to the view that the learner does indeed select a single grammar, and that it is the MAP grammar. This is an assumption of convenience.

Bayesian inference provides only a computational-level specification of the problem of learning (or perception, or selecting actions, etc), in the sense of Marr 1982. Bayes’ Rule says what should be computed as the best solution to a given problem—but not anything about how. It stands to reason that this is all we could conclude: Bayes’ Rule simply states an equality as a static fact. Criticisms have arisen of the extreme computational lengths one needs to go to to do even the most coarse and approximate inference using certain types of interesting Bayesian models (Stevens 2011, Berwick, Pietroski, Yankama & Chomsky 2011). These criticisms are misleading, because they place a demand on the Bayesian models that they were never intended to fulfil. The reductionist reasoning that motivates the use of Bayesian models is that human behavior/learning/inference is deducible from general principles of optimization—that is, just from the statement of the

problem (including the way that the learner represents its input from the outside world, the learner’s “intake”: Gagliardi 2012) plus Bayes’ Rule is sufficient to determine what learners will do (modulo the questions just raised).

Now, it is true that, if it is obvious that the full optimization required is intractable, then we risk raising more questions than we answer. Nevertheless, to carry out Marr’s goal of explaining cognition, one must investigate both of the computational and the algorithmic levels, and then go on to connect them. The top-down approach asks whether human behavior in, say, learning falls out as an “optimal solution” to the problem of balancing a “degree of faithfulness to the data” with a “lack of markedness of the solution”; the bottom-up approach specifies an algorithm and asks whether it matches human behavior. However, such an algorithm will generally be consistent with some implied assessment of fit and relative biases. To the extent that, for example, perception can independently seen as implying the same fit-to-data function, and to the extent that the biases in learning reflect preferences that can be seen in non-learning tasks, we also need to explain why it is that the learning algorithm appears to optimally balance these two independently motivated forces; and, if it does not, why it does not. The most interesting cases are the ones where the optimal solution appears to be incorrect with respect to human behavior (Pearl, Goldwater & Steyvers 2010, Gagliardi, Bennett, Lidz & Feldman 2012, Stevens, Trueswell, Yang & Gleitman 2013, Gagliardi & Lidz In press)—but we cannot show this without knowing what the optimal solution would be.

Accusations of unfalsifiability have been levelled at the Bayesian program because it occupies the computational level (most recently by Jones & Love 2011). The best response to this is essentially what we just said: since prior distributions and decision rules



in cognitive applications constitute empirical hypotheses, they should ideally make some independent predictions. However, another (top-down) approach to constraining the problem would be to develop a theory of “natural” or “automatic” prior distributions following in some way from the structure of the problem; while it is well known that there is no universal criterion for picking the unique “natural” prior distribution for a given problem, (van Fraassen 1989), independent of how the problem is parameterized, I will make a suggestion towards some weak constraints of this kind in grammatical inference in Chapter 2.

This is enough about Bayesian inference for now. As the reader has already been encouraged to imagine, it has the potential to be a powerful tool for linguistic theorizing, because the goal of linguistic theory is to explain how it is that human beings come to knowledge of language from primary linguistic data. Although linguists are very good at formulating hypotheses about what adult states are possible and impossible, we do not very often fill in the crucial blanks about the relation between the data and the final state (Viau & Lidz 2011, Pearl & Lidz 2009). Chapters 2–5 are intended as an illustration of how to make good on the tantalizing promise that Bayesian inference can step in to fill in these blanks in a principled way.

## 1.2 Phonology

Human language is a complicated system. However it is that human beings actually turn out to work, we as researchers at least generally decompose the system into several different mappings. One mapping, the semantics, can be broadly construed as exchanging

information from the whole host of cognitive systems that furnish us with our “thoughts” (vision, event tracking, theory of mind, and perhaps a system of “general” concepts) with structured “meaning” representations, with help of one kind or another from pragmatics, the system by which one reasons about the intentions of the interlocutor. The semantic representation is often thought to be exchanged with a representation of more abstract “dependencies” by the syntax, a mapping with its own principles, whose representations are then exchangeable for either an abstract perceptual representation of an utterance with that particular meaning, and/or instructions for using the motor systems to generate one. This latter mapping is called phonology.

Phonological information is organized into minimal units called morphemes; a scaffolding for these elements, or for slots for these elements, is provided by the syntactic structure, which is often thought to be a tree. The morphological structure, that structure which holds within a narrower and still incompletely understood domain called a “word,” which organizes morphemes but which itself sits in the syntactic scaffolding, is sometimes thought to follow different principles from the syntax (and sometimes not, in which case syntax is said to deal in morphosyntactic structures). Morphemes associate the information in the syntactic representation and the corresponding semantic representation with arbitrary “chunks of pronunciation.” Whatever the “meaning” information is that is associated with the concept dog is associated with the essential information key to recognizing or producing the word dog. Some morphemes are also thought to be associated with parts of the syntactic structure that are “strictly grammatical,” and which bear only some indirect relation to the semantic representation, such the English passive marker, which is often thought to work by playing some formal tricks forcing noun phrases into one

position rather than another (associated with a pair of morphemes we write be+en, as in John was eaten by a lion). The morphemes, at any rate, along with their morphosyntactic scaffolding, are where our story begins.

The phonological computation is usually divided into two parts, the phonological grammar and the phonetics–phonology interface. Morphemes, on the morphosyntactic end, are finite sequences of elements called segments. The other two ends could be thought of as sequences, too: under a simple view of phonetics, the perceptual representation is just a sequence giving the perceptual system a map with which to recognize each segment as it comes in: d-o-g, and so on; and, under such a simple view, the production systems are also furnished with finite sequences of motor instructions. The job of the phonetics–phonology interface is to exchange these perceptual or production representations for the types of sequences found in the morphemes themselves. The phonological grammar, on the other hand, does some manipulations which actually modify the content of the representations.

In English, for example, the word alternation seems to be alternate+ion, or something like this, in terms of its morphological decomposition. However, the pronunciation of alternate is with a final [t], but the corresponding sound in alternation is [ɪ]. This is a general pattern: deletion, elation, emendation, and so on; it also seems to be related to the pattern [d]/[ɪ] found in corrode/corrosion, erode/erosion, and so on. This pattern is called an alternation if we want to focus on the generalization that sometimes [ɪ] appears apparently in place of [t] ([ɪ] “alternates with” [t]), and it is said by way of explanation that the phonological grammar executes a rule which changes the [t] to [ɪ] in some particular environment, or that the phonological grammar at least gives rise to a process by which [t] changes to [ɪ] in that environment.

Similarly, in Hungarian, we find that “our windows” is *ablak+unk*, [ɒlkunk] and “our glucoses” is *glukóz+unk*, [lukozunk], but “our cauldrons” is *üst+ünk*, [ytynk], and “our chauffeurs” is *sofőr+ünk*, [oførynk]. The alternation is between [u] and the front rounded vowel [y]. In fact, it is a much more general alternation, whereby (simplified slightly) suffix vowels mutate into other vowels that “correspond” in some sense to the last vowel in the stem. (They need to match that vowel in the front vowel/back vowel dimension, but everything else about them stays fixed: [u] corresponds to [y] in the back vowels because they are both high; but here we see the alternation being triggered by the non-high vowels [], [o], and [ø].) This is a process called Vowel Harmony.

In Spanish, we see [b] alternating with the fricative [β], as in *vámos*, [ɓamos], *cabra*, [kaβra], as well as [d] alternating with the fricative [ð], as in *dámelo*, [damelo], *cada*, [kaða]. A simple version of the generalization for this rule (Spirantization) is that stops change to fricatives after vowels. Finally, in English, the Deaspiration alternation is virtually always the first rule that is given to linguistics students: *top*, [t<sup>h</sup>p], *stop*, [stp]. The puff of air (aspiration) which is ordinarily pronounced after [t], [p], and [k] seems to disappear when [s] appears immediately following in the same syllable.

At least some of these alternations are morphologically active—they actually show evidence, via their morphological composition, that one particular segment changed into another one. These are cases where we can also do wug tests (Berko 1958) and find that speakers really do change things on command: we make up a word like *wug* and we ask English speakers to inflect it in the plural form. We get [wz] for the plural of *wug*, but [wks] for the plural of *wuck*, following another general rule of English (Voicing Assimilation). Patterns like English Deaspiration are also often treated as active processes by way of

having a particular hypothesis about how they arise and what speakers know about them, but they do not show direct evidence of a “change” like this (see Chapter 3 for discussion of some relevant speech perception data for cases like this, however).

A still more important point to stress about these alternations is that they seem to show categoricity. This fact is fundamental to phonology. The idea is that, in memory, the phonetic details are not (all) stored with a segment; the phonemes in the lexicon pick out a subset of the set of all phonetically possible segments. We evidently need to make more fine-grained distinctions than this when we are specifying exactly what the motor systems should do, or what the perceptual systems should attend to (if we could not do this then we would not be able to explain all the minute differences in pronunciation that we see from one language to the next). This is why we say that the phonetics–phonology interface engages in a translation, to fill in this gradient detail. Apart from categoricity, the other non-trivial property of lexical representations is that they are segmented. This just means that the categories are defined over rather coarse chunks temporally (consonant or vowel-width chunks). The motivation for this is as simple as that: there are some temporal chunks within which the phonetic information falls into equivalence classes; none of the narrower temporal distinctions seem to need to be made in order to state phonological processes. There is also some evidence for segments from speech errors (see Chapter 3) and resyllabification (see Chapter 4); these patterns not only show that the chunks have fairly substantial width, but also that the chunks are smaller than syllables. A third property of lexical representations is that they are decomposed into cross-classifying features, but I will save the discussion of this for Chapter 3.

Processes for which the result is phonetically identical to or indistinguishable from

the pronunciation of some other phoneme, or from the output of some other process—are called neutralizing processes.<sup>2</sup> (Strict) allophony is another type of pattern, which provides a weaker kind of evidence for categoricity: the results of the processes do not correspond phonetically to any other segment, but the results are at least internally consistent. In Spanish, the [β] and [ð] sounds only appear as the result of Spirantization, (with the result that they only occur in one, very restricted set of contexts), but it seems that [β]’s are at least fairly phonetically similar to each other when they are pronounced, and do not just appear as pronunciations of totally random sounds, or track finite details of the the pronunciation of the preceding triggering vowel (at least this latter point is usually assumed, but the issue has not been thoroughly investigated). Cases where a segment only arises due to some process, but two different underlying segments can generate it, are generally called allophony too, because they generate segments which are (by hypothesis) not found in the lexicon. Another term for this is non-structure-preserving. These phenomena will be the subject of much discussion in this dissertation.

Finally, a few details are in order about phonological theory. The classical derivational theory of phonological grammar was first fully articulated in *The Sound Pattern of English* (SPE; Chomsky & Halle 1968). A sample computation for the (simplified) Hungarian Vowel Harmony pattern is shown here:


---

<sup>2</sup>Although in Chapter 5 I cite two examples of published experimental studies which find results like this—total phonetic neutralization by a phonological alternation, into the same equivalence class—and another in passing in Chapter 3, the results are easy enough to recreate using phonetic corpora, which are now widely available. The interested reader is invited to create graphs like the ones I provide for the “incomplete” neutralization cases in Chapter 5, only for the English alternation alternation (try the Buckeye corpus: Pitt, Johnson, Hume, Kiesling & Raymond 2005) and Turkish Vowel Harmony (use the METU corpus: Salor, Pellom, Ciloglu & Demirekler 2007). The result will contrast sharply with the graphs in Chapter 5.

			lukoz + ynk
(12)	$V \rightarrow [\alpha \text{ back}]$	$\left[ \begin{array}{c} V \\ \alpha \text{ back} \end{array} \right]$	$C_0 + C_0 - \text{ lukoz + unk}$
			lukoz + unk

The top row shows the underlying lexical representation; the bottom row shows the output of the phonological grammar (only explicitly specified in the “forward” direction). The notation in the left column shows the rule that executes Vowel Harmony: vowels change to match in the feature [back] with a preceding vowel when zero or more consonants, then a morpheme boundary, then zero or more consonants, intervene. The dash indicates the position of the target segment. Two crucial properties of this theory are (i) that the grammar can contain arbitrarily many rules, and they compose in some way that is also specified as part of the grammar (here only one is shown); (ii) the learner’s inference is over these individual rules. This contrasts sharply with Optimality Theory (OT; Prince & Smolensky 2004). A tableau representing an OT computation is shown here (analysis from Ringen & Vago 1998):

	$\begin{array}{c} /luk\ o\ z\ y\ nk/ \\   \quad   \\ +b\ -b \end{array}$	Ident-IO <sub>st</sub> [bk]	Align-R[bk]	Ident-IO[bk]
(13)	$\begin{array}{c} [luk\ o\ zynk] \\   \ / \\ +b \end{array}$			*
	$\begin{array}{c} [luk\ o\ z\ y\ nk] \\   \quad   \\ +b\ -b \end{array}$		*!	
	$\begin{array}{c} [luk\ o\ zynk] \\   \ / \\ -b \end{array}$	*!		*
	$\begin{array}{c} [luk\ o\ z\ y\ nk] \\   \quad   \\ +b\ +b \end{array}$		*!	*
	$\begin{array}{c} [luk\ o\ z\ y\ nk] \\   \quad   \\ -b\ -b \end{array}$	*!	*	**

This theory works by generating every output (actually, every possible input–output mapping) for a given input, then filtering these until only one remains. Only a few crucial candidates are shown in any given tableau. There is a universal set of constraints which assign violations (\* marks) to candidates. These constraints are given a language-specific ranking (highest-to-lowest: read left-to-right). The computation descends down the list of constraints and only the best candidates—the ones with the fewest violations of the current constraint—are allowed to pass. The rest are excluded (an ! is marked after the first fatal \*). Only the best survives (marked with ). This analysis also assumes autosegmental theory (which is a theory of how segments are represented and of the basic operations of grammar that is orthogonal to OT; Goldsmith 1976). In that theory, segments can share features; these features are usually shown like this, dangling from their associated segments, to indicate that they are independent of each other in various ways (see Chapters 4–5 for brief discussion). The relevant constraints shown in this tableau are: Ident(ity)-I(nput)O(utput), in two flavors: one restricted to the st(em), and another



working across the board. This type of constraint examines the correspondence between segments in the underlying representation and their yields in the output. In this case, the value of the b(ac)k feature needs to be identical, and a violation is assigned for each instance where this is not true. The Align-R(ight edge) constraint, keyed to the [back] feature again, assigns a violation for each instance of a vowel feature intervening between the right edge of the rightmost [back] feature and the right edge of the word.

The key features of OT are that (i) for the purposes of the grammatical computation, there is only one compositional “step” (it is monostratal); the constraints only get to make reference to one single input representation (the original) and one single output representation (the final one) throughout the process of computation; (ii) the basic elements of the grammar are not the actual steps taken to modify an input but the constraints on what modifications are allowed—or, rather, a ranking of these constraints; and (iii) these constraints are “output-oriented,” and never make reference to just the input (although input–output correspondence constraints are fine). Some points to watch for below are: (i) the grammar should actually have at least two steps, because some of what used to be in the phonological grammar will move to the phonetics–phonology interface (somewhat reminiscent of Stratal OT). This is found in Chapters 3–5. For (ii), it has never really been clear what exactly the “representation of the grammar” (in this case, as filters versus instructions to make changes) is supposed to correspond to cognitively. If it computes the same function as some other type of grammar, what is the difference? In this case one could perhaps point to some difference between performing many operations to generate many candidates, and performing one operation to generate one candidate, but it is never clear how many of the intermediate data structures in a grammatical computation one is supposed to

take seriously, or how exactly they are supposed to stand in correspondence with exactly what brain-instantiated features of the computation. If there is no such “data structure” interpretation, then all that remains is (ii): the grammar is stated in one way rather than another. I suggest that the only agreed-upon locus for this “structure of a grammar” is in acquisition; in the Bayesian framework, this structure must somehow be reflected in the prior probability distribution on grammars. This is found in Chapter 2. Finally, for (iii), I propose that, in the phonetics–phonology interface component, grammatical statements are actually crucially input-oriented in a way that does not comport with either OT or SPE-type theories.

It is also worth noting briefly that OT analyses generally take non-alternating cases of allophony not to actually be allophony, but rather just knowledge of a combinatorial restriction on segments—although this not an absolutely crucial part of the theory. They can afford to do this with some elegance, because the alternation and the statement of the static generalization do not need to be stated separately: they both arise out of the filtering process. This is fine in principle: different descriptive generalizations can be drawn out of the same finite data set, neither being more “empirical” than the other, and they lead to different causal theories. However, the description of allophony, as I will argue in this dissertation, has always been far more contingent than just a simple debate over whether a pattern constitutes a “static” or “active” alternation. Implicitly, all the standard analyses assume a transcription has been provided, which neatly classifies things into allophones; but, as any beginning field worker or introductory phonetics student knows from their difficulties in coming up with a classification for phones, this is a tremendous and risky simplification. The peril of this classical approach for the learner is the subject of Chap-

ter 3, and the theory—that allophones are in fact a part of a phonetic module far more respectable than it is usually given credit for—carries through to the end. This concludes my brief overview of the basic material of this dissertation.

## Chapter 2: Simplicity

Mystery. You're always surrounded by them. But if you tried to solve them all, you'd never get the machine fixed.

—Robert Pirsig, *Zen and the Art of Motorcycle Maintenance*

### 2.1 The poverty of the stimulus: what is to be done?

This chapter develops the theory of language acquisition. To outline how arguments about language acquisition go, and to set the stage for a particular proposal about acquisition, let us begin with a family of informal arguments about acquisition called “poverty of the stimulus” arguments. From Chomsky 1975:

Suppose ... the child has learned to form such questions as [is the man tall?, is the book on the table?, etc.], corresponding to the associated declaratives [the man is tall, the book is on the table]. ... [T]he scientist might arrive at the following tentative hypothesis as to what the child is doing ...:

Hypothesis 1: The child processes the declarative sentence from its first word (i.e., from “left to right”), continuing until he reaches the first occurrence of the word “is” (or others like it: “may,” “will,”

etc.); he then preposes this occurrence of “is,” producing the corresponding question ....

This hypothesis ... is false, as we learn from such examples as [the man who is tall is in the room—is the man who is tall in the room?, the man who is tall is in the room—\*is the man who tall is in the room?]. ... Children make many mistakes in language learning, but never mistakes such as [\*is the man who tall is in the room?]. ... The correct hypothesis is the following ...:

Hypothesis 2: The child analyzes the declarative sentence into abstract phrases; he then locates the first occurrence of “is” (etc.) that follows the first noun phrase; he then preposes this occurrence of “is,” forming the corresponding question.

... Hypothesis 2 holds that the child is employing a “structure-dependent rule,” a rule that involves analysis into words and phrases, and the property “earliest” defined on sequences of words analyzed into abstract phrases. ... [T]he scientist must ask why it is that the child unerringly makes use of the structure-dependent rule postulated in hypothesis 2, rather than the simpler structure-independent rule of hypothesis 1. ... A person may go through a considerable part of his life without ever facing relevant evidence, but he will have no hesitation in using the structure-dependent rule, even if all of his experience is consistent with hypothesis 1. The only reasonable conclusion is that UG contains the principle that all such rules must be structure-dependent.

(31–32)

A more recent description of this particular case—cleaned up slightly to avoid the misleading procedural metaphors about grammatical knowledge—is to be found in Berwick, Pietroski, Yankama & Chomsky 2011: the interpretation of auxiliary verb-initial sentences like *Can eagles that fly eat?* obeys a “structure-dependence” restriction. This restriction has the effect of ruling out “is it the case that eagles that can fly eat?” as a possible interpretation, evidently because the initial auxiliary verb must always be interpreted as modifying the nearest verb which is at the same structural “level,” (that is, Can [eagles that fly] eat), and never a closer verb with respect to the linear order of items in the string, but which is at a lower structural level (Can [eagles that fly] eat). The relevant principle might be stated in this case as “interpret the fronted auxiliary as modifying the nearest verb at the same level of bracketing in the grammatical structure”—call it Principle S (corresponding to Hypothesis 2). The pieces of the argument from learning go as follows:

- (14) Here are some logically possible hypotheses the learner could entertain (about how a linguistic system might work)
- (15) Here is the input available to an inference mechanism operating over these hypotheses
- (16) Conclude: If the actual learning outcome is one thing, but something else was logically possible, then evidently that logically possible hypothesis is not actually possible

Much of the controversy surrounding these arguments—and the counterarguments—stems from the fact that it is rare for either to contain all the logically necessary details. Here are what the pieces should be:

- (17) Delimit a full range of possible hypotheses
- (18) Say what the input is
- (19) Specify the behavior of an inference mechanism
- (20) Conclude: (assuming the specification of the input is correct) If the actual learning outcome differs from the prediction, the proposed set of hypotheses or the inference mechanism must be different from what was proposed

More specifically, if we conclude that the proposed inference mechanism would be unable to decide between the actually correct hypothesis and some other, or would arrive at some incorrect hypothesis, then one obvious move is to propose that the set of hypotheses is actually restricted and does not include the incorrect hypothesis. Apart from that, this all seems quite clear, but it is also clear that the path through a debate about any of this is fraught with danger if we mistake the simplified argument in (14)–(16) for a fully worked out argument like (17)–(20).

For example, one type of debate has looked at the facts about the input and the learning outcomes. Do children really not hear enough examples of sentences like *Is the man who is in the room tall* and *Can eagles that fly eat* to prefer Hypothesis 2 over Hypothesis 1? And do they really not show evidence of ever preferring Hypothesis 1? (Crain & Nakayama 1987; Sampson 2002; Legate & Yang 2002.) But all this empirical work is beside the point if it only addresses these two possible hypotheses. As Lasnik & Uriagereka 2002 point out, whatever mistakes learners turn out to make or never make, whatever type of evidence turns out to be absent, we will probably still be able to adduce vastly or infinitely many contradictory ways for them to characterize it which they seem to

consistently avoid. Consider some hypothetical alternatives to Principle S, adapted from Lasnik and Uriagereka:

- (21) Interpret a fronted auxiliary as modifying any verb
- (22) Interpret a fronted auxiliary as modifying the first verb that comes after a complete phrase-structural constituent

Both of these new alternatives are consistent with all the data consistent with Chomsky's Hypothesis 1, but also with a lot of data which would be good enough to rule it out (like hearing and understanding *Can eagles that fly eat?*). However, both are still radically incorrect, as there is in fact never any ambiguity as to the verb being modified in these cases, and because there are still more complex sentences (such as *Can eagles that fly and swallows that sing eat?*) which would be incorrectly interpreted by the second principle.<sup>1</sup>

The leap from a single example of a wrong hypothesis to a larger set of vastly or infinitely many logical possibilities is actually left implicit in many of the classic arguments about induction, not just this one from Chomsky: Goodman 1955 probes our intuitions as to whether a scientist or other rational observer would feel that a set of green emeralds, all observed before time  $t$ , could ever confirm the hypothesis that all emeralds are “grue,” where “grue” means “green if examined before time  $t$  and blue otherwise”; we think not, of course, and he says that this just goes to show that “lawlikeness” is a non-trivial concept in need of further investigation—and that such an understanding is furthermore crucial to understanding induction. Similarly, Quine 1960 concludes that correct translation is im-

---

<sup>1</sup>Note that the second principle makes crucial reference to phrase structure but is still wrong: although the principle playing the role of Principle S in versions of this argument is often given the convenient label of “structure-dependence,” that label is by no means sufficient to characterize the principle—what is proposed is that the interpretation of sentences makes use of structure in a very particular way, namely, Principle S.



possible in many cases, after imagining a linguist probing a consultant for the meaning of the utterance “gavagai” in the presence of a rabbit. The linguist could not decide on the basis of an affirmation by the consultant of the appropriateness of the utterance in the presence of any number of other rabbits whether it referred to a rabbit or a collection of undetached parts of a rabbit, or a collection of temporal slices of a rabbit (the last two both include actual rabbits as particular cases), and so on. Even if the linguist were to find the right words to ask, he would face similar problems with the meanings of those words, and so on. In each case the argument made to convince the reader of the existence of a problem of induction is presented by way of one or two particular examples, followed (implicitly or explicitly) with “and so on.” It is fair to say that it is really the “and so on” on which such informal arguments rest, for better or for worse: this discrepancy between (14) and (17) makes a huge difference.

The absence of (19) in the simplified version is also quite important. But something needs to be said about this before we can say anything at all. The only allusion to a mechanism that connects hypotheses and data in the quote from Chomsky is the allusion to the relative “simplicity” of Hypothesis 1: if there is some measure of simplicity of hypotheses that guides the learner, then we can use that to make predictions about what ought to be learned given a particular set of data and hypotheses. However, once some details of the inference mechanism are spelled out (more than this), we realize that that mechanism, and not only the set of alternative hypotheses, is also a possible candidate for explaining the discrepancy between the “logically possible” and the “actually learned.”

Within linguistics, much of the reason that (14)–(16) has subbed in for (17)–(20) has been the understanding that, once the set of hypotheses was correctly specified, nothing

more needed to be said about the inference mechanism—that it was trivial. This idea was developed in the 1980s in generative grammar, largely following the publication of Chomsky 1981. Pursuing a conception of language that implies that the set of “core” syntactic grammars is actually finite, (what became known as the “Principles and Parameters,” or “P&P,” approach), Chomsky argued that “it is quite possible” that one could restrict attention to a proscribed subset,  $S$ , of bounded-length sentences, and then find, for each possible grammar, “decision procedure ... that enables the  $n$  grammars to be differentiated in  $S$ ” (11). A decision procedure in this context is simply a function that returns true or false for any given sentence in  $S$ , telling us whether the sentence is permitted under grammar  $i$ . The idea of “differentiating the grammars in  $S$ ” reduces, in this context, to the further conjecture that  $S$  exists such that no two grammars share exactly the same extension, restricted to  $S$ . Then, given a large enough subset of  $S$ , eventually we can identify the grammar.

This is an argument against immediately ruling out as “unlearnable” any grammar for whose language a decision procedure does not exist (e.g., a grammar generating all and only positive instances of the Halting Problem, the Post Correspondence Problem, etc). The idea is that the hardness of the decision problem associated with a particular grammar says nothing about how hard it would be to successfully learn grammars from some set which happens to contain that grammar. The fallacious conflation of these two different kinds of hardness is based on the mistaken idea that the inclusion in the hypothesis space of a grammar that is hard in the decision sense immediately implies that all members of the “usual” class to which that grammar is considered to belong—in this case, the whole set of recursively enumerable functions, which is to say, any function, computable or not—is therefore included in the hypothesis space too. Obviously, if the problem space for the

learner was, “pick some grammar, about which I have told you nothing except that it is some function you might be able to dream up,” the learner would be faced with a very hard problem! But the division of functions (grammars) into classes for the purposes of analyzing their computability or complexity, although illuminating, is not equal to the division we care about for learning, namely, which grammars is the learner able to learn, and which grammars is the learner not able to learn. That is the point, and this is a good argument for that point.<sup>2</sup>

However, from this point on in LGB, Chomsky seems to largely take for granted the point that, given that the set of grammars is finite, a practical and trivial learning procedure exists that can recover the grammar from data. The logic just outlined is not a good argument for that point. In fact, it is not an argument for that point at all. The procedure whose existence is conjectured (non-constructively) is a counterexample provided against one particular fallacious line of reasoning, as just outlined—nothing more. Subsequent years of research in P&P confirmed that, indeed, the problem of language acquisition is in no way trivial just because the set of grammars is finite (Dresher & Kaye 1990; Niyogi & Berwick 1996; Yang 2002; Pearl 2007).

---

<sup>2</sup> Heinz (2013) reiterates this point quite clearly, but, in spite of this, the misapprehension persists in the field that the argument goes further, and makes irrelevant all and any use of classes of functions established in mathematics or computer science, including the Chomsky hierarchy, in arguments about the human language faculty. This is incorrect; many well-established facts with sweeping empirical implications for the theory of human language, such as the fact that phonological grammars are uniformly sub-regular while some syntactic grammars fall outside the context-free class (Kaplan & Kay 1994, Shieber 1985, Heinz & Idsardi 2011) would not even be stateable without relying on this body of research. As with anything in linguistics, a result will be interesting to the extent that it helps us narrow in on the question of how exactly knowledge of language is acquired, and what that constitutes, a key part of which is narrowing in on the range of possible grammars using whatever tools. Chomsky’s argument can be paraphrased as follows: understanding the distinction between recursive and recursively enumerable functions is not sufficient to solve the problem of language acquisition; the incorrect interpretation of the argument is, at its most pernicious, that analyzing grammars in terms of classes of functions well-established in mathematics is not necessary to solve the problem of language acquisition—an obvious delusion.

Thus there is no escaping the fact that the familiar “poverty of the stimulus” argument, (14)–(16), is not a substituted for a fully worked out argument of the form (17)–(20). What is to be done? Fill in the details. There are plenty of interesting things to be learned about the set of actually possible hypotheses, and about the inference mechanism. However, in order to draw any conclusions about what they are, we must first state clearly what exactly the properties are that are at issue. An informal argument predicated on the very narrow property “contains one very particular hypothesis X” might help convince us that neither the set of hypotheses nor the inference mechanism turn out to be trivial or obvious. But it would be misguided to keep pursuing that question as a way to convince a skeptic of whatever stripe, because that property is neither interesting nor easy to reach firm conclusions about without filling in virtually all the details about the inference mechanism. We need to state precisely what it is that we are really trying to prove or disprove about the set of actually possible hypotheses for the learner, and firmly nail down enough assumptions about the inference mechanism to allow us to reason about whether those properties hold or not; similarly, *mutatis mutandis*, for any property of the inference mechanism—it needs to be general but precisely stated, and we need to be able to fix some assumptions about the set of hypotheses, before we can draw any conclusions about it.

In this chapter I discuss the inference mechanisms; I will say things about the set of hypotheses only in the interest of making a point about these inference mechanisms. The ones I will focus on in this chapter are any which comply with Bayesian inference; I will argue that such mechanisms are indeed a source of explanation. The goal is to establish a particular “law of inference” which holds in a Bayesian learner, preferring intuitively “simpler” hypotheses (the Bayesian Occam’s Razor). In particular, the goal will be to

tighten up previous accounts of this phenomenon to lay out, given a Bayesian inference mechanism, how to characterize the sets of possible hypotheses and data sets that will lead to such a preference. As Bayesian inference is a class of mechanisms with the prior distribution as a free parameter, I also give a condition on this prior that will guarantee that we get the law to hold (I introduce something called the Framework Consistency Principle that priors need to follow). I finally show that there is a way in which we can see the application of this principle for generating prior distributions as the application of a plausibly domain general “optimal inference” principle to acquisition, and I spell out what this would mean for learning grammars for human languages.

In the next section I give a brief introduction to Bayesian inference as a language acquisition mechanism, and review one recent and prominent debate in the literature involving Bayesian inference but which in my view fails to sufficiently engage with certain relevant issues crucially tied to Bayesian inference, namely, the Bayesian Occam’s Razor; the rest of the chapter is in a sense intended to fill in these important and interesting details in that argument.

## 2.2 Bayesian models of cognition: why should we care?

Bayesian models of cognition have flourished in recent years. Bayesian models of cognition use probabilities to track relative “preferences” for different cognitive states—in the case of language acquisition, for different possible grammars. They are specified in two parts: a likelihood function, a mapping from grammar–data pairs to probabilities, representing a “fit” score assigned to the data by the grammar; and a prior distribution,

a mapping from grammars to probabilities, representing the a priori preferences of the learner. The posterior distribution is a mapping from grammar–data pairs to probabilities, representing the learner’s scoring of a particular grammar as a model of a given data set. The posterior distribution can be derived mechanically as a function of the prior distribution and the likelihood function, which is a key benefit of Bayesian inference with broad-reaching practical implications for constructing complex models. Computationally intensive methods for stochastic search are often used to obtain a representative sample of relatively high-posterior grammars when the exact grammars of interest cannot be practically specified for comparison in advance; however, with few exceptions, it is the posterior, and not these purely instrumental algorithms, which constitute the model (see Pearl, Goldwater & Steyvers 2010, Phillips & Pearl 2012 for attempts at constraining inference in psychologically meaningful ways). In abstracting away from the search procedure to focus on relative preferences, Bayesian models take what is in linguistics called the evaluation measure research strategy: do not specify how the learner works, but merely specify the “learning relation,” from  $\text{input} \times \text{internals} \rightarrow \text{learning outcomes}$ , as it would apply under some idealized circumstances (Chomsky 1964, Chomsky 1965, Chomsky & Halle 1965, Chomsky & Halle 1968).

The evaluation measure approach to linguistic theory traditionally applied a function from grammars (not grammar–data pairs) to values or costs which would, minimally, allow the learner to select a unique high-valued grammar in case two grammars were equally consistent with the data. The received understanding was that this was the only case that the evaluation measure would need to handle, but this rested on the assumption that “consistency with the data” was a binary distinction (fully consistent or fully inconsistent), an

obvious simplification (Chomsky & Halle 1968: an exception is LSLT, Chomsky 1975, which attempted the more ambitious goal). This simplification was in part a consequence of the common simplification of the notion of grammaticality to a two-way distinction: if an utterance is observed, and it is assumed for the purposes of learning that what is observed is necessarily grammatical, then the appropriate notion of “consistent with the data” is obviously “predicts that all of the observed data is grammatical,” and “inconsistent,” conversely, “predicts that at least some of the observed data is ungrammatical.” Making the assumption that all of the individual observations must be jointly consistent in order to have any degree of consistency with the data at all is also crucial to maintaining the binary notion of consistency. Thus the binary distinction of consistency is clearly a simplifying assumption. That it is a simplification suggests that we might want the evaluation measure to bear on the learner’s behaviour not only when the degree of consistency with a given data set is equal across two hypotheses, but also when it is different.

The Bayesian approach to language acquisition asserts that the consistency function (which takes as input grammar–data pairs) and the evaluation measure (which takes as input simply grammars) are both probability measures, conventionally called the likelihood function (probabilities over data) and the prior measure respectively (probabilities over grammars); probability measures output real numbers between zero and one, where zero is a minimum and one is a maximum for both consistency with the data and value under the evaluation measure. Crucially, it also asserts that (almost) all that needs to be known about the learner’s final state given some data can be summarized by a probability measure as well, called the posterior measure, (giving probabilities over grammars) which can be

derived from the prior and likelihood as follows:

$$(23) \quad \text{Posterior}[G, X] \propto \text{Likelihood}[G, X] \cdot \text{Prior}[G]$$

In (23),  $G$  represents a grammar and  $X$  represents some observed data. Although the constraint imposed by (23) alone specifies a family of measures proportional to each other, the assumption of a unit measure (a measure that integrates to one) means that the posterior is unique (the scaling constant must always be the reciprocal of the integral of  $\text{Likelihood}[\cdot, X]$  over the set of grammars  $\mathcal{G}$  with respect to the measure  $\text{Prior}[\cdot]$ ). If the posterior is thought of as the conditional probability of a grammar  $G$  given data  $X$  and the likelihood as the conditional probability of data given grammar, then (23) amounts to an application of Bayes' Rule (conditional probability will be discussed in more detail below). In the context of language acquisition, we may call the derived function the posterior evaluation measure, and, when the distinction is crucial, I will call what is traditionally referred to as simply the evaluation measure the prior evaluation measure.

The idea is that the posterior evaluation measure is sufficient to guide the choices of the learner. If we believe that the learner eventually stops learning and selects a single grammar, then this might be the highest posterior valued grammar; on the other hand, if we believe that the adult state is one where there is still some lingering uncertainty as to what the correct grammar for the ambient language is (as suggested, for example, by the approach of Yang 2002), then the full posterior evaluation measure is needed to characterize it.<sup>3</sup>

---

<sup>3</sup>The wrinkle responsible for the “almost” above is that, in the broader Bayesian setting, posterior probability measures are usually combined with additional apparatus (a loss function and a decision rule) which



Much recent research in language acquisition has made use of the assumption that learning outcomes can be understood as falling out from Bayesian reasoning about grammars. This research strategy is useful because it reduces the burden on the next researcher to come along—the one who needs to specify the actual algorithm and/or implementation by which the learner actually comes to the right grammar—reducing it to the problems of computing the posterior distribution from the bias and goodness-of-fit functions (basically trivial using Bayes’ Rule) and optimizing over that function—a massive gain, as evidenced by the diverse range of optimization technology and criteria that have been thrown at the language acquisition problem (Dresher & Kaye 1990, Clark & Roberts 1993, Niyogi & Berwick 1996, Yang 2002). Even without doing any search, or attempting to examine the full posterior, it can be very illuminating to look at the shape of the posterior distribution: if the grammar(s) that matches human behavior is worse, according to the posterior distribution, than ones that do not, then, either the theory of grammar (i.e. the combination of bias/prior and grammaticality/likelihood functions), or else the Minimalist assumption that the learner selects the “best” grammar in the rational sense (that is, following the evaluation measure that follows from its bias and grammaticality functions), is incorrect.

Apart from being an interesting empirical conjecture, this assumption has several practical benefits. Two of these were raised directly in the exchange between Perfors, Tenenbaum & Regier 2011 (henceforth PTR) and Berwick, Pietroski, Yankama & Chomsky 2011 (henceforth BPYC). First, Bayesian statistical approaches do not burden the

---

complete the characterization of the behavior of the actor, (in this case the learner), of which the posterior measure represents only a part (the internal belief state). Like most other cognitive scientists, I make the assumption that we can just think of some trivial apparatus here, with the understood caveat that working out the full decision-making apparatus is not optional in the end if we believe a single grammar is “selected.” See Chapter 1.

researcher with irrelevant representational choices in the way that, for example, connectionist models do: neural network models demand real-vector valued inputs, outputs, and internal representations, which requires that researchers make substantive choices in order to convert back and forth between these representations and symbolic representations that are really already at the limits of our detailed knowledge about the mental encoding of language. This conversion can sometimes be awkward (as in the case of Rumelhart & McClelland 1986's famous Wickelphone representation) and can sometimes represent a major research undertaking in and of itself (for example, the representation of hierarchical structure in neural network models still has no agreed-upon general solution). In contrast, the Bayesian approach allows the researcher to simply program whatever data structures seem appropriate as representations, because it is a method for associating numbers (posterior values) with these structures and nothing more. In this way, researchers can avoid unintended consequences of what are usually arbitrary choices. Second, Bayesian statistical approaches to inference obey various general and interesting laws, which the Bayesian approach to cognition inherits as laws of reasoning.

For PTR, both the easing of the burden of representational choice, and the fact that certain laws of inference hold, are factors that make the Bayesian approach useful in studying language acquisition. They present a Bayesian evaluation measure for syntactic grammatical knowledge which they apply to real (but simplified) child-directed utterances. On the first issue, they write, in their conclusion, that:

[W]e have offered a positive and plausible “in principle” alternative to the negative “in principle” poverty-of-stimulus arguments for innate knowledge

of hierarchical phrase structure in syntax. ... By working with sophisticated statistical inference mechanisms that can operate over structured representations of knowledge such as generative grammars, ... we can more rigorously explore a relatively uncharted region of the theoretical landscape: the possibility that genuinely structured knowledge can be genuinely learned, as opposed to the classic positions of nativism (structured but unlearned knowledge) or empiricism (learned but unstructured knowledge, where apparent structure is merely implicit or emergent). (331–332)

In other words, although learning arguments yield conclusions about inductive biases, the issue has been dealt with as if it were also about what types of representations the mind uses—the question of structured representations, and related questions—even when those inductive biases could be cashed out either way. This conflation is at least in part an artefact of the types of acquisition models used previously, so, using Bayesian methods, one can now more easily refute a claim about the necessity of some strong inductive bias without smuggling in irrelevant representational claims. Such an example makes it clearer that learned/not-learned is a question that is orthogonal to the question of structured/not structured.

On the issue of laws of inference, they allude to one particular law of inference that comes out in many Bayesian systems, writing that:

These results emerge because an ideal learner must trade off simplicity and goodness-of-fit in evaluating hypotheses. The notion that inductive learning should be constrained by a preference for simplicity is widely shared among

scientists, philosophers of science, and linguists. ... The tradeoff between simplicity and goodness-of-fit can be understood in domain-general terms.

(313)

They are referring to the Bayesian Occam's Razor. The idea that there is something "domain-general" about the prior evaluation measure PTR present comes up repeatedly in their paper, and, although it is clear that the issue of domain-specificity is supposed to be somehow tied up in the learning argument being made, it is never clear precisely how. As discussed above, there are two basic approaches to explaining facts about acquisition: adjust the assumptions about the hypotheses available to the learner, and adjust the assumptions about the inference mechanism. The Bayesian Occam's Razor is a fact about the inference mechanism (it says that a large class of prior measures are forced into being biased toward simpler hypotheses by virtue of the basic facts of probability theory being embedded in the inference mechanism). On the other hand, a choice of possible hypotheses is implicit in any choice of prior (or indeed in any discussion of learning). One or both of these things is a plausible candidate for which part of their model is meant to be "domain-general." As I argue below, although PTR do not say (or believe) it, the use of Bayes Rule and, more interestingly, the Bayesian Occam's Razor embedded in the prior, are really the only plausible candidates for something "domain-general" about their acquisition model.

Before proceeding, however, it is important to stress what the limits of this type of research are and in particular how it relates to "innateness." The conclusions above, in (16) and (20), are about something inside the human mind and that is the only sense in

which they can be said to tell us anything about “innateness.” In particular, for some authors (including, but by no means limited to, PTR) the term “innate,” as applied to (either) a set of possible hypotheses or an inference mechanism, means or is at least perfectly correlated with the property “domain-specific.” This sense of “innate” (mechanisms specific to language) goes far beyond the property “biologically fixed” and even farther beyond the property “mind-internal,” two reasonable alternate meanings for this massively overloaded term. But the conclusion that the set of possible internal systems that are possible outcomes of learning (the “actually possible hypotheses”) is restricted in some particular way cannot by itself tell us anything about whether that restriction is an idiosyncrasy of language or vision or whatever system the learning is taking place in; similarly for conclusions about the inference mechanism (which, as pointed out by Pearl & Lidz (2009), must be assessed separately for its domain-generalty or domain-specificity; similarly for the encoding in which the input comes in, which is distinct from the “grammar” in the narrow sense). To reach such further conclusions, one would need to examine different cognitive systems and compare them, and then hypothesize a relation between the two systems that says in what way they are related. Arguments of the form (17)–(20) do not do this, and more generally neither does any argument which simply starts with some assumptions and assesses whether they support acquisition. We will need to keep this in mind as we review PTR’s argument and BPYC’s reply in the next section: although PTR say they are doing something having to do with assessing the domain-specificity of some representational capacity in syntax, this is wrong; they are not.

## 2.3 The syntactic acquisition model of Perfors et al.

PTR present a prior for simple syntactic grammars and compute the posterior for some simplified corpus data (strings of part-of-speech tags representing utterances from CHILDES: MacWhinney 2000). The prior is over context-free grammars generating strings of these tags, and the question is whether the learner will prefer context-free grammars which are in or out of the set of right-regular (strictly right-branching) grammars — actually is a bit more complicated than this, but we will leave it at this for the moment.

This is interesting, according to PTR, because the properly context-free grammars have true “hierarchical structure,” whereas the right-regular grammars, the subset of the context-free grammars which are strictly right-branching, are not much more than glorified lists (clarification for the syntactician: the c-command relation is equal to the precedence relation, except for the bottom node; clarification for the computer programmer: working with Lisp should make it clear in what sense linked-lists are “right-branching structures”; obligatory note for the computer scientist rusty on formal language theory: right-regular grammars generate regular sets, hence the name). They show that the induced posterior measure does indeed favor context-free grammars over right regular grammars.

PTR claim that this research is related to the classic example poverty of the stimulus argument given above, having to do with the structure-dependence of syntactic rules. BPYC question this claim. In this section, I will review BPYC’s argument for why this is incorrect, and elaborate on their discussion of why flaws in PTR’s logic mean that the paper is actually irrelevant to any interesting questions about how restrictive or unrestricted the set of possible syntactic grammars could be, and still support learning. I will also go

over again why, even if they did have something to say about this, it would be irrelevant to questions about the domain-specificity of syntactic mechanisms. Finally, I will point to the fact that there actually is something extremely important and interesting about PTR's paper, but it is buried; this turns out to be the Bayesian Occam's Razor, which is the topic of this chapter.

Let us begin with strictly right-branching grammars. As BPYC point out, PTR's terminology is misleading: the regular grammars they use to stand in for "non-hierarchical" structure in fact yield hierarchical structures which happen to be strictly right-branching. They are not the same, intensionally, as grammars which generate strings using flat structures, although there is a conversion between the structures yielded. The issue goes beyond terminology, and it goes beyond potentially subtle issues of internal mental representation. The way that almost any linguistic theory of phrase structure works, one is always forced to let in strictly right-branching structures as possible analyses for at least some sentences (it is difficult to find a modern syntax paper without a strictly right-branching analysis for at least one sentence). More importantly, the way that almost any theory of syntax using phrase structure works, one is always forced to let in grammars that could only ever generate right-branching structures! This is very problematic for PTR. The implication in their paper is that "hierarchical"—i.e., for them, not strictly right-branching—grammars are "really" language-like, and "non-hierarchical"—i.e., strictly right-branching—grammars are not. This, for them, is evidently why it is important to ask whether it can be learned that the best grammar for a given corpus is "hierarchical." But, in fact, strictly right-branching grammars are perfectly language-like, in that they are predicted to be possible in most syntactic theories.

Of course, it is difficult to imagine how a learner could assign a strictly right-branching structure to sentences like *The man likes John*, where the man is obviously a complex constituent, a noun phrase—let alone prefer a grammar that gives a strictly right-branching analysis to every possible sentence. The fact that it is difficult to imagine should suggest why it is not particularly surprising that the learner is able to rule these analyses out, but it is important to see that such a “defective” analysis is possible, because this is exactly the sort of analysis of the corpus that PTR’s learner comes to rule out (whether that is as significant as they claim or not).

Some informal reasoning about how acquisition might take place will make clear what this kind of strange acquired grammar would look like. Intuitively, one might think that a learner hearing only the sentences *The man likes John* and *John likes the man* would be compelled, minimally, to at least the following conclusions: since *John* and *the man* both occur after *likes*, but *John likes the* never occurs, the constraint on what follows *likes* must treat *John* and *the man* as of a kind and say that this class of things (of “constituents”) can follow *likes*, as opposed to asserting that *the* can follow *likes* and *man* can follow *the* independently; the learner has independent evidence about the fact that *the man* is of a kind with *John* for the purposes of word order restrictions because *Man likes John* never occurs, and the reasoning follows in a parallel way; so the only method the learner has to analyze *the man* is using some (learned) principle of analysis into noun-phrases. To see the (informal) reasoning more clearly, label noun phrases *J* (for “John-type”). Given the sort of syntactic representation we are talking about, this means the learner comes to believe in something isomorphic to the simple  $J \rightarrow \text{the man} | \text{John}, S \rightarrow J L, L \rightarrow \text{likes } J$ , which is not strictly right-branching. This is our intuition about what should happen, an



analysis so obvious that it is hard to see that it is not the only one.

Actually, there are a lot of assumptions in this informal supposition which do not follow from any formal constraints on grammars. For example, it is not logically necessary that the learner analyze all instances of John or the man in the same way; the learner might instead adopt the right-branching grammar  $J \rightarrow \text{the man} | \text{John}$ ,  $S \rightarrow \text{John } L | \text{the } M$ ,  $M \rightarrow \text{man } L$ ,  $L \rightarrow \text{likes } J$ . John likes the man is then analyzed as  $_S[\text{John } _L[\text{likes } _J[\text{the man}]]]$ , and The man likes John as  $_S[\text{the } _M[\text{man } _L[\text{likes } _J[\text{John}]]]]$ .

Such an analysis is possible even under the “merge” theory presented by BPYC, which is a basic version of what has come to be called “minimalist” syntactic theory (Chomsky 1995), even though there are substantial restrictions on phrase structure implied in that theory. For example, the theory asserts that syntactic constituents give a privileged status to one of their elements (the “head”), and that there is a label assigned to the whole constituent which is uniquely determined by the head. All combinatorial restrictions are then stated in terms of these constituent labels. In terms of the sorts of rules we have outlined here, this implies that if there is a rule  $L \rightarrow \text{likes } J$  and a rule  $? \rightarrow \text{likes}$ , then  $?$  must either be  $L$ , or else there must be two accidentally homophonous lexical items likes. But absent constraints from the semantics, it is easy to make up a right-branching grammar that generates our little corpus that is subject to these restrictions:  $S \rightarrow \text{the } M | \text{John} | \text{John } L$ ,  $M \rightarrow \text{man } L$ ,  $L \rightarrow \text{likes } S$  (that is, “man” has label  $M$ , “likes” has label  $L$  and “the” and “John” have label  $S$ ). Although this seems implausible on the face of it, and of course generates even more unattested sentences such as John and The man likes John likes John, the point is that this analysis cannot be ruled out by formal constraints under this theory, or many others like it; the explanation for why must have to do with the inference

mechanism, then (perhaps something to do with the fact that the grammar overgenerates).

If we introduce enough accidental homophony, we can even do better by the over-generation problem: if the John in The man likes John is different from the John in John likes the man, then we are free to assign them different labels. With such a powerful tool as homophony in hand, it is easy to see how the learner could come up with a (perverse) strictly right branching grammar for just about any corpus—and with homophony, as with the kinds of funny (mis-)labelling required to get the previous grammar to work out, there cannot be an outright ban, so if a grammar like this is not the grammar learners arrive at, then there must be some reason other than a ban, a soft preference—hidden somewhere in what we call the posterior evaluation measure. In sum, right-regular grammars are not impossible or non-syntax-like, and no one ever said they were; the idea that it is somehow contrary to standard linguistic theory to posit an acquisition device that can learn to use “hierarchical” grammars, choosing them over these supposedly “unstructured” analyses, is incorrect. Such an acquisition device is, rather, taken for granted.

Before explaining why this also has nothing to do with the Principle S issue discussed above, we should clean up the characterization of the PTR prior, which has been somewhat unfaithful. What PTR claim to be doing is not merely learning a grammar for a corpus. In their words, “the interesting claim ... is not about the rule [or grammar] for producing interrogatives ( $G$ ) per se; rather, it concerns some more abstract knowledge  $T$ . ...  $T$  is the knowledge that linguistic rules are defined over hierarchical phrase structures. This knowledge constrains the specific rules of grammar that children may posit and therefore licenses the inference to  $G$ ” (311, underlining mine). This is actually a very tricky idea, and it may become clear to the reader only after the discussion of model evaluation

below. The way it works is like this: the prior is actually not just over grammars—this is the slight inaccuracy alluded to above. Actually, what is learned is a grammar–indicator pair,  $\langle G, \omega \rangle$ :  $G$  is the grammar, and  $\omega$  is a single bit, whose values I will for convenience map to  $\{r, c\}$ . The role of  $\omega$  is that, if  $\omega = r$ , then  $G$  is restricted to being right-regular; when  $\omega = c$ , it is not restricted in this way, it is only restricted to being context-free. Equivalently, the acquiendum  $\langle G, \omega \rangle$ , where  $G$  is not right-regular, is impossible if  $\omega = r$ ; if  $G$  is right-regular, then  $\omega$  may be either  $r$  or  $c$ . The inference about  $\omega$ , then is a “higher-order” inference, in the sense that what we (the learner) take its value to be actually determines how we do inference about  $G$ . We will see in the discussion of model evaluation below that such inferences about what kind of solution we are allowed to infer are useful in practical settings, such as selecting which predictor variables should or should not be included in the analysis of some experimental data we might collect. We will also see, throughout the chapter, that these sorts of nested, or hierarchical, inference problems permeate the problem of language acquisition.

To take just one example, consider the problem of learning words: when listening to an utterance, the learner may believe that utterance consists entirely of known words, or may decide that it actually contains a new, unknown word. If the utterance seems to be composed entirely of known words, then the learner may (or may not) take the opportunity to adjust how it thinks the words in its lexicon are pronounced, or what they mean, etc., but the range of in-principle-possible lexicons is just the same as it was before; but if the learner believes that a new word has been uttered, then a whole new range of possibilities are opened up, most notably the possibility that the contents of the lexicon—the pronunciations, semantics, and grammatical information—are just exactly what the learner already

thought they were, except that there is a new lexical item, corresponding to a new tuple of pronunciation/semantic/grammatical information. Thus, conditional on some inference that the learner makes, the learner’s range of possible “solutions” to the problem “what are the contents of the lexicon” changes. Going back to the PTR prior, we might characterize what is being learned as simultaneously learning about what grammar is ambient and about what grammar is.

Hierarchical inferences are always difficult to interpret.<sup>4</sup> For a moment, the idea of “learning what grammar is” might seem to be meaningful or even profound; but, upon reflection, it is actually quite difficult to understand. Put aside the idea that “discovering” the class of grammars, rather than merely discovering the grammar, is somehow a more “domain-general” inference: we already dispensed above with the notion that learning arguments of this kind can say anything about domain-specificity.

To see what kind of question PTR are asking, consider a more general case. Suppose that  $\omega$  actually ranges over the whole non-enumerable set of subsets of the set of all rewrite grammars of every kind, so that one value of  $\omega$  picks out the right-regular grammars, another picks out the context-free grammars, another picks out the context-sensitive grammars, another picks out the (Turing-complete) set of all unrestricted rewrite grammars (Type 0 on the Chomsky hierarchy), another picks out the strange assortment of grammars that would be obtained by constructing some particular mapping between real numbers and grammars and then picking only those grammars paired with prime

---

<sup>4</sup>This is why the discussion of simplicity in Bayesian inference continues to be at an impasse in the philosophical literature: although the exegesis would take me too far afield, Forster & Sober 1994, Dowe, Gardner & Oppy 2007, and Henderson, Goodman, Tenenbaum & Woodward 2010 are at cross-purposes and fail to engage with some of the crucial conceptual issues, which ultimately have to do with the hierarchical Bayesian inferences that give rise to the Bayesian Occam’s Razor I discuss in this chapter.

numbers—the reader gets the idea. Now contrast this with a prior that lacks  $\omega$ , and only needs to pick out some grammar  $G$  from the set of all (unrestricted) rewrite systems. The range of possible grammars (i.e., functions) that the learner can acquire is exactly the same. The potential “models” of any observed data are the same, although one might say that the “explanation” can vary under the pair–learner, in a sense that is somewhat difficult to make more precise. Short of spelling out this rather thorny intuition, what can we say is the motivation for the higher-level inference in this case?

For example, imagine that the usual arguments from linguistic theory about restrictions on grammars (or, perhaps more neutrally, “internal language perception/production models”) are incorrect: there are no hard restrictions on possible grammars; rather, any logically possible mapping is possible. This is what many “emergent” explanations of language amount to, or aspire to amount to—a claim there is no “innate knowledge” constraining the hypothesis space (for language, although again, domain-specificity is actually orthogonal). Put the intension of the mapping aside—the claim is merely that the net result of learning “could be anything.” What does this have to do with whether the learner simultaneously infers a restriction on the class of mappings it is selecting from? We might expect this to be the case if the learner were trying to determine whether what it was hearing was, say, an animal call ( $\omega = a$ ) or human language ( $\omega = l$ ), and then had some differing biases about what the concomitant structure would be in either case. Or, supposing one were presented with data from many languages, asked to learn grammars for all of them and then to form some generalization about those grammars, then there would be an obvious motivation for adding a variable corresponding to “class of grammars”—although, as this is not the situation the language learner is put in, we would

surely interpret the higher-level inference as one that a scientist, and not an actual language learner, would do. But, in any case, without some empirical or conceptual reason for assigning grammars/analyses to the various different classes corresponding to different value of  $\omega$ , the idea that grammatical inference has this additional structure is merely an interesting conjecture, not motivated by anything. More importantly, it is no more or less like any set of standard assumptions in linguistics or anywhere else to assume, or not assume, that the set of right-regular grammars is specially delimited and elevated to some special status distinguishing it from the set of more general context-free grammars. As we will see, the real difference between a learner with and without the additional  $\omega$  inference is that the learner with the additional inference believes that right-regular grammars are simpler in a sense that the learner without it does not. It is just not clear what the viability of such a learner has to do with any pertinent issues in cognitive science.

It remains to explain why PTR are wrong that inference under their prior bears on the Principle S issue discussed above—and they say that it does. Citing Chomsky on the example of subject–aux inversion, they write that “only by defining syntactic rules ... over hierarchical phrase structure representations is a child likely to be able understand that [Is the boy who is smiling happy?] expresses a certain complex thought while [\*Is the boy who smiling is happy?] expresses no well-formed thought. Hence our focus here is on the more basic question of how a learner can come to know that language should be represented in terms of hierarchical phrase structure” (310). I have just argued that, (a), following BPYC, “hierarchical or not” actually is not well-aligned with the distinction (right-regular or not) that PTR try to map it to, (b), again following BPYC, given that the set of possible grammars that the PTR prior allows is roughly that which would be

posited under any linguistic theory, the “coming to know that language should be represented in terms of hierarchical phrase structure” reduces to the rather more underwhelming prospect of “learning a grammar,” and (c), adding an additional inference about the class grammars are drawn from in a situation where there is only a single grammar at issue is barely different from “learning a grammar,” except that it adds a subtle claim about the “structure” of the prior—an idea which we will see come up throughout the chapter, but which has no clear cognitive interpretation in this case. Even given these points, however, there might be some interest if the questions being addressed bore indirectly on a question, structure-dependence of so-called  $\bar{A}$ -dependencies, which has been the subject of a reasonable amount of previous research (Crain & Nakayama 1987, Legate & Yang 2002). However, although PTR cite such an argument as the main motivation for their study—correctly noting that the availability of a tree-like representation is a prerequisite for any grammatical computations depending on that representation—the mere availability of such representations says nothing about how they will be used in computation. “Structure-dependent” is shorthand for “structure-dependent in a very particular sense”—namely, Principle S—a sense which does not in any way follow from the existence of hierarchical structure.

Finally, I return to the question about domain-specificity. PTR write that “[BPYC] have argued that much recent work ... misses the original intention of the argument. ... Our goal here is to explore what we see as a basic issue at the heart of language acquisition—the origins of hierarchical phrase structure in syntactic representation,” with the question about “origins” evidently being over whether hierarchical phrase structure “comes from” language or not: “the argument about innateness is primarily about the role

of domain-specificity,” (307). Innateness has nothing to do with domain-specificity. This is very important. One could indeed imagine the PTR study coming out differently, with the learner always favoring right-regular grammars, even when they were inappropriate, thus suggesting an “innate” (inborn, biological, natural) bias in favor of them; or, one might imagine that a learner with the higher-order inference over grammar classes is especially drawn to the conclusion that the grammar should be drawn from the narrower class, the right-regular grammars (we will see below why one might have expected such a thing to happen in this model). Given that these are evidently not the grammars people actually come to, this would suggest a model of the mind in which there is no such higher-order inference. In either case, there is something to be concluded about the “innate” structure of the mind. There is nothing, however, to be concluded about whether that particular innate structure is special to language or is shared when humans learn in some other domain. One way to ask that question would be to apply the same model, or a model with comparable structure, to data from another domain. It is unfortunate that the word “innate,” whose familiar meaning seems easy enough to apply to the relevant issues in cognitive science, has come to mean “innate and domain-specific” when applied by cognitive scientists, an obviously detrimental conflation. There is no way to draw any conclusions here about the domain-specificity of the set of hypotheses.

Nevertheless, I will argue that there is something plausibly domain-general about the prior, in particular, the automatic Bayesian Occam’s Razor it conceals. Although it would require further research to establish parallels in inference across domains, this is at least a plausible candidate for a domain-general mechanism: it falls out under the combination of Bayesian inference (which has nothing to do with language per se) and a way of asso-



ciating hypotheses (grammars) with prior distributions that could be thought of as being “minimalist” in Chomsky’s (1995) sense of “following from the structure of the problem in some optimal way.” It would not be surprising, therefore, if under close scrutiny this law of inference were found in other domains and were found to hold for the same reasons.

We begin with the necessary background about evaluation measures and about Bayesian statistical theory.

## 2.4 Evaluation measures: restrictiveness and simplicity

As discussed above, an evaluation measure for grammars is a function mapping from grammars to some comparable values, intended as a description of the learner’s relative preferences. Evaluation measures were a subject of interest relatively early in the history of generative grammar. As there will be multiple grammars equally consistent with a given data set even under a restrictive theory, the evaluation measure is motivated as an additional device to constrain the behavior of the learner.<sup>5</sup>

For example, Chomsky & Halle 1965 presented the following two candidate de-

---

<sup>5</sup>Evaluation measures have sometimes also been referred to as “evaluation procedures” (in certain ambiguous passages in Chomsky 1965 “evaluation measure” and “evaluation procedure” seem to be conflated) or “evaluation metrics” (Bach & Harms 1972). I avoid the former term since it is extremely misleading, the crucial property of evaluation measures being that they imply nothing about any actual procedure; I avoid the latter term, although it is the dominant one, to avoid the misleading implication that evaluation measures are, or need to somehow induce, metrics in the mathematical sense (that is, distance functions); the term “evaluation measure” suggests that evaluation measures need to obey the axioms of a measure, which may not be true in general, but which is acceptable in the current context, as the evaluation measures developed here are indeed measures in the mathematical sense.

scriptions of a set of phonological facts in Western Mono (taken from Lamb 1964):

$$(24) \quad \begin{array}{l} w \rightarrow k^w / \quad h- \\ k^w \rightarrow q^w / \quad V_1 h - V_2 \end{array}$$

$$(25) \quad \begin{array}{l} w \rightarrow \left\{ \begin{array}{l} q^w / \quad V_1 h - V_2 \\ k^w / \quad h- \end{array} \right\} \\ k^w \rightarrow q^w / \quad V_1 h - V_2 \end{array}$$

In (24) and (25),  $V_1$  and  $V_2$  are two different classes of vowels. When flanked by these vowels, the underlying sequences /hw/ and /hk<sup>w</sup>/ are both realized as [hq<sup>w</sup>]; otherwise, they are both realized as [hk<sup>w</sup>]. If the grammars, crucially the one in (24), are understood as rule systems ordered as given, then both grammars characterize this pattern. In (25), the grammar has been modified to avoid making crucial use of rule ordering. However, presumably, the learner must decide on one representation, and yet both must be available.

The evaluation measure proposed by Chomsky & Halle 1968 is that “the ‘value’ of a sequence of rules is the reciprocal of the number of symbols in the minimal schema that expands to this sequence” (334). This would lead the learner to grammar (24) in the face of two equally consistent grammars. In general, the evaluation measure was supposed to trade off against a graded evaluation of fit, but specifying the fit quantitatively was ignored in practice.<sup>6</sup> As discussed above, this evaluation measure corresponds to a prior, and not

---

<sup>6</sup>In Chomsky and Halle’s words: “We will not concern ourselves here with the nontrivial problem of what it means to say that ... a proposed grammar ... is compatible with the data ... . In other words, we make the simplifying and counter-to-fact assumption that all of the primary linguistic data must be accounted for by the grammar and that all must be accepted as ‘correct’; we do not here consider the question of deviation

a posterior, evaluation measure in a Bayesian setting.

This evaluation measure is intuitively guided by “simplicity”: grammars which are “simpler” in terms of the number of theoretical objects they contain are more strongly preferred. Apart from simplicity, the other main intuition guiding evaluation measures has been restrictiveness. To take another phonological example, Hale & Reiss 2008 consider two proposed versions of a rule from Georgian, which has the standard five vowel inventory {i, e, o, u, }, but no low front vowel [æ]:

$$(26) \quad l \rightarrow l \ / \ - \left[ \begin{array}{l} + \text{ ATR} \\ - \text{ back} \\ - \text{ low} \\ - \text{ round} \end{array} \right]$$

$$(27) \quad l \rightarrow l \ / \ - \left[ \begin{array}{l} + \text{ ATR} \\ - \text{ back} \\ - \text{ round} \end{array} \right]$$

In Georgian, underlying /l/ is realized as [l] before [i] and [e]. Unlike the Western Mono case discussed above, these two grammars are not extensionally equivalent; (27) predicts that the rule should apply to all [−back] vowels, while (26) predicts that it will apply to only [−back, −low] vowels. Thus (27) maps /læ/ to [læ], while (26) maps /læ/ to [læ]. However, given that Georgian has only one [+low] vowel, (and assuming that it is treated as [+back]), both grammars will be equally consistent with the evidence. Contrary from grammaticalness, in its many diverse aspects,” (Chomsky & Halle 1968, 331).

to the symbol-counting evaluation measure, which would value (27) more highly, Hale & Reiss 2008 propose what amounts to the opposite: “The correct statement of a rule ... is the most highly specified representation that subsumes all positive instances of the rule, and subsumes no negative instances of the rule” (103). This principle (which would form the basis for a posterior evaluation measure, in our sense) makes reference to only two possible values, correct and, implicitly, incorrect; it apparently assigns correct if no specification could be added to the rule without making it inconsistent with the data, otherwise incorrect. Although it might appear to be stated, like Chomsky and Halle’s evaluation measure, in terms of notation, Hale and Reiss make it clear that the motivation is not notational complexity per se, but, rather, restrictiveness: “The more specific, that is, more restrictive, rule is the one provided by the [language acquisition device]” (104). Here, “more specific” refers to the amount of information in the statement of the environment, where each piece of information is treated as a constraint, or restriction; a more specific statement will necessarily be more restrictive. In this case, the notion of restrictiveness aligns with increased notational complexity.

Simplicity and restrictiveness together form the basis for most of the general principles of language acquisition proposed in the literature. They are sometimes at odds, as in the Georgian example (but not in the Western Mono example). In spite of this, they can be combined, as indeed Hale and Reiss seem to do implicitly when they assume that learners form a single rule handling both /le/ and /li/, rather than formulating the generalization as two rules rather than one. (This “minimal generalization” proposal can be found throughout the literature on phonological acquisition: Pinker & Prince 1988, Yip & Sussman 1997, Albright & Hayes 1999, Dunbar 2008).

The Principles and Parameters approach to linguistic theory focused attention on restrictiveness, in the form of the Subset Principle. One reason for this is simply an idiosyncrasy of the facts that P&P focused on. In the case of an obligatory phonological rule, as in the Georgian example above, when restrictiveness is invoked to decide between competing environments, it is not because the set of possible surface strings, or the set of possible underlying–surface mappings, is more restricted under one grammar than another; both these sets will always have one element for each possible underlying form, and will merely change depending on the restrictiveness of the environment for a given rule. On the other hand, the prototypical case to which restrictiveness was applied in P&P involved allowing or barring the optional null subject *pro*. Allowing it has the result of strictly expanding the set of strings (and of sound–meaning pairs). It is presumably because of this difference that researchers generally agreed that the learner’s response should be to value the more restrictive grammar (the one that generates the subset, rather than the superset): if the alternative strategy is to always prefer the less restrictive grammar, then this strategy will surely fail, as both grammars will be equally consistent with evidence coming from the subset grammar.

Another reason for the focus on restrictiveness is that it is unclear what simplicity would mean in the context of P&P. Unlike in rule-based theories, in P&P theory, the core parts of a grammar are represented as a fixed-length sequence of parameter values; being of fixed length, no notion of notational simplicity is available. More recent family-related syntactic theories have reduced “parameter setting” to selection of lexical items, and the lexicon is of arbitrary length no matter what the theory; but in the case of adding a lexical item which has not been attested (such as *pro*, given data from the subset language),

intuitive simplicity (fewer lexical items) coincides with restrictiveness.

Optimality Theory is another linguistic theory in which the principal issue in constraining acquisition has been taken to be enforcing restrictiveness. There is always a possible analysis of any phonological system in which the lexicon is simply a record of the surface forms, and the grammar is trivial; but real grammars are thought to generalize somewhat. The analysis needs to be restrictive in order to prevent the grammar from being general in a way that generates impossible surface forms for items not in the lexicon. This subset problem arises under any theory, but it attracted attention in OT because of the “Richness of the Base” theory, under which the learner cannot posit the language-specific morpheme structure constraints (restrictions on possible lexical items) which are often appealed to to block some of these cases of overgeneration. Simplicity was once again sidelined, for the same reason as in P&P theory: target OT grammars consist of a total order over the universal set of constraints; thus, on the face of it, every grammar has the same size (the cardinality of the universal constraint set), and notational simplicity has no obvious place in the grammar.

As previously alluded to, however, simplicity is only displaced in these theories, not irrelevant. In both cases, the lexicon must be learned, and the choice of a set of stored forms, and thus the choice of a more or less compact set of stored forms, has profound consequences for the rest of the grammar simply because the two parts of the analysis depend upon each other.

In syntactic theory which has developed since P&P, parameter settings are understood as lexical choices (Chomsky 1995), and the acquisition of the lexicon is therefore the only kind of learning there is. For example, Bobaljik & Jonas 1996 reduce a complex

set of facts about Germanic languages to a single parametric difference in the possibility of the Tense head having a specifier position (differences in the availability of object shift, verb raising, and “transitive expletive” sentences with both an expletive and an overt subject, such as the grammatical Icelandic equivalent of *There have many Christmas trolls eaten pudding*). For Bobaljik and Jonas, this parameter is encoded as the presence or absence of a strong Determiner (that is, NP movement-inducing) feature on the Tense head. If this grammatical choice is treated as the presence or absence in the lexicon of a Tense head with a strong D feature, or as the optional presence or absence (via underspecification) of such a feature, then it is without question a simpler grammar that does not allow a strong Determiner feature on Tense. Since choices about the contents of the lexicon are all that can vary across grammars in this theory, simplicity could hardly be more relevant.

Similarly, although Hale and Reiss argue for restrictive learning of phonological alternations, there are many cases in which generalization beyond the data obviously happens, a choice which under many theories would constitute a notationally simpler grammar. For example, English speakers extend the voicing alternation seen in [læfs] versus [dz] to nonnative (thus unseen) segments such as [x], as in [bxs] (*Bachs*). One possible account of such facts is to simply rule out the more restrictive grammar a priori: if there is no constraint in the universal constraint set that can mark disagreement in voicing for [f], [], and so on, as more problematic than disagreement in voicing for [x], then there is no way of encoding the unattested subset grammar, and the choice disappears. A more involved solution is proposed by Hale and Reiss, who similarly claim that universal feature theory rules out a single feature matrix specifying an environment with [f] and [] but not [x]. They, however, must rule out the possibility of specifying such an environment using a disjunc-

tion of feature matrices, the availability of which is generally understood to be necessary for descriptive adequacy in rule-based theory, and they rule this out on the grounds that the learner collapses such disjunctions into a single feature matrix wherever possible (see above). This suggests that these disjunctive environments are excluded or dispreferred on what are intuitively simplicity-driven grounds. Empirical evidence supporting overgeneralization in phonology, whether in early acquisition or in historical change, has often been explained by appealing to a bias toward simpler grammars (Kiparsky 1971, Bach & Harms 1972; see also Goro 2007 for a case of apparent overgeneralization in scope acquisition in Japanese).<sup>7</sup>

Even in Optimality Theory, there has always been a suggestion that the learner has the capacity to expand the set of applicable constraints in some way, and recent computational models have attempted to induce phonotactic constraints within OT-type grammatical frameworks, in particular, Hayes & Wilson 2008, Adriaans & Kager 2010. As soon as the grammar becomes variable in length in this way, questions about the value of simplicity become quite relevant, and indeed both of these constraint induction systems rank constraints for the purposes of acquisition by their generality, which happens to be quite similar in effect to notational simplicity given the way these learners compute it.

Finally, although overt appeals to simplicity biases are somewhat out of fashion, implicit appeals to simplicity biases are still ubiquitous in linguistic theory: virtually every analysis of any pattern will be guided by the principle that it is better to collapse two

---

<sup>7</sup>Morphological examples suggesting historical development in the opposite direction, towards greater notational complexity, were cited by Kiparsky 1985; the interesting speculation that these cases might be explained by learners systematically weighting certain parts of the primary linguistic data over others was put forth by Lahiri & Dresher 1984, who presented evidence that learners sometimes pay special attention to certain forms over others in driving grammatical changes. It would be interesting to develop this idea quantitatively in terms of the influence of the likelihood function.



patterns into one wherever possible. For example, Zimmermann 2002 gives the following completely run-of-the-mill description of some facts about German and English: in some languages, distance distributives, like the “each” in The boys have bought two sausages each, or “jeweils” in the German translation Die Jungen haben jeweils zwei Würstchen gekauft, can distribute only over individuals; in others, they can distribute over events. This refers to the following two interpretations for the sausage sentences:

(28) “Each of the boys has bought two sausages”

(29) “The boys have bought two sausages each time” (say, each time that they went to the butcher)

The first interpretation is always available, but the other one is possible in some languages (like German) and impossible in others (like English and Dutch). Zimmermann’s thesis is entirely devoted to an account of these distance distributives, and centers around the issue of how it is that *jeweils* can be interpreted in two different ways in German (but not in English and Dutch). However, such an analysis, and, in fact, even the descriptive generalization formulated above, presupposes that *jeweils* is the same lexical item in both cases; if the two meanings are actually due to two different words that are simply accidentally homophonous, then we already have an answer to this question. How to analyze the two obviously family-related meanings is still an interesting question, but there ceases to be any demand for a single meaning common to both.

This kind of assumption, that words that look the same are the same, is a fundamental tenet of linguistic analysis; any time there is a way to account for the same facts using an “uninteresting” appeal to an alternate, differently-behaving lexical item, most linguists

will choose to do otherwise. This is why we choose to account for phonological patterns rather than adding idiosyncratic lexical markings each time we observe a morpheme behaving differently, or, even more radically still, abandoning morphological analysis entirely (although it so happens that simple wug-test experiments tell us that grammars also choose to account for phonological patterns). The idea is that positing one lexical item is a better analysis than positing two.

It is crucial to understand that this is not an application of Occam's Razor, which says to the analyst that "theoretical entities are not to be multiplied beyond necessity." Occam's Razor, in its normal interpretation, is a principle for the scientist, which guides the construction of theories. It says that simpler scientific theories are better. For example, Jackendoff 1977 proposed to do away with general phrase structure rules of the form  $A \rightarrow B \cdots C$ , because there actually seemed to be systematic relations between the labels on parent nodes and their children (an *NP* will always contain an *N* and so on). Jackendoff's revision,  $\bar{X}$ -theory, caught on, presumably because it was a simpler theory: fewer phrase structure rules are possible under that theory than were under the previous theory.

But the tenet of "no homophony," and the more general tenet of "fewer lexical entries," is not a guide towards simpler linguistic theories; it is a guide towards simpler linguistic analyses, and alternate analyses are not alternate theories about how the linguistic system works, but rather conjectures about of which grammar a learner has posited, and a grammar is a "theory" of a different sort (the internal mental model that has been "theorized" by the learner to account for the language). When we say that one analysis is impossible or implausible because it is more complex, then we are attributing a choice to the learner; it is not our choice to make, but rather an empirical hypothesis about what

choices the language acquisition device makes. The ubiquity of such reasoning in linguistic theory represents an implicit scientific claim that demands justification.

## 2.5 Bayesian Occam's Razor

### 2.5.1 Maximum likelihood and restrictiveness

This section is about statistics. It ultimately introduces a law of inference called the Bayesian Occam's Razor, which has to do with simplicity, but in this subsection a different law is presented, which has to do with restrictiveness. The goal is to properly understand both the similarities and the differences between these two inference effects once both have been presented. To relieve the burden of the long explanations of statistics that follow, I begin with an example. Suppose we are given a collection of observations, presented here as a histogram. Which of the three curves in Figure 2.1 is a better model for this data?

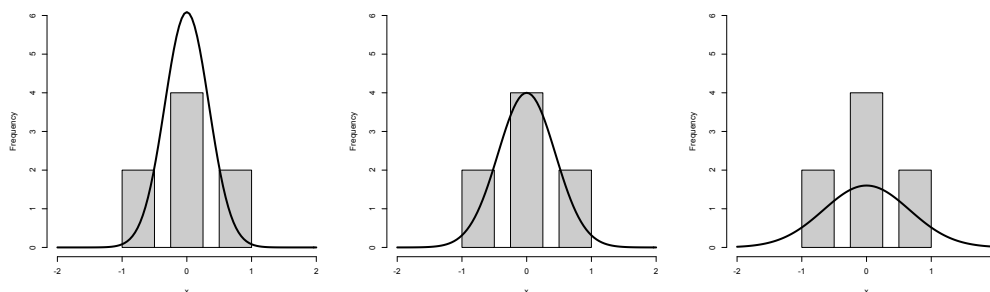


Figure 2.1: Three Gaussian models for some data.

The one in the middle is. Why? First we must make sure we have some familiarity with what these curves mean: they are in a sense pictures of normal or Gaussian probability

distributions, and in particular they are density functions,<sup>8</sup> which means that the height of the curve is directly related to how probable each of these three models says that various different observations are. All three curves predict that the actually observed data are more probable than the data that was not observed—this is good. But the curve on the left is so tall in the middle that the observations to the two sides are predicted to be somewhat less probable than under the second curve; and the curve on the right is so broad that the observations in the middle are predicted to be a lot less probable than under the second curve. The second curve is just right.

This choice follows from something called the maximum likelihood principle. The maximum likelihood principle implies that absence of evidence is evidence of absence. The reason is scarcity. We only have a finite amount of credence, and so we must cling to it and not give it away for free to observations that never showed up—and we do not get points for putting our credence in possible observations, only for actual observations. Although this sounds silly, it is exactly why the maximum likelihood principle says that the curve in the middle is the best. In particular, the likelihood under each of these three hypotheses is the probability of the data given that hypothesis,  $\Pr[X|H]$ .<sup>9</sup> The rule for evaluating the likelihood given a collection of data points in a case like this is that we multiply together the probabilities that we get back from the likelihood for each of the

---

<sup>8</sup>Except that they have been rescaled. This does not matter—the alignment of the outputs of the density function with the absolute number of observations matters not at all for any evaluation; only the relative values matter. But this rescaling makes the curves look like nice “models” for the data, on the grounds that the pictures look the same.

<sup>9</sup>This is simplifying slightly, which will continue to be the case as we continue to ignore the usual use of the density function to compute likelihoods in the continuous case, and not the associated probability measure. For this statement to be actually true, we would need to be assuming that the observations are actually observations of some narrow range of the real line, not single points. Points have probability zero. As this is probably the sensible thing to assume if we think about it long enough anyway, and after all this is what the picture shows, (although it is obviously not what it is supposed to mean), assume this.

data points. So, the higher the individual probabilities according to the curve, the higher the likelihood. But, since it is a probability measure, the likelihood function, integrated over the whole real number line of possible observations, needs to integrate to one: the area under the curve needs to be one. (For presentational purposes, these curves are not properly scaled to integrate to one, but they do all integrate to the same constant.) This is because the area under the curve in any region of the number line represents the probability of such an observation, and the area under the whole curve represents the probability of any number at all, which is of course the maximum probability—one—because if we see a number, it must fall into the range “any number at all.” So, if we add some height from the curve in one place, then we must remove some elsewhere to get the area, or “probability mass,” to balance out along the number line. Adding to the middle takes away from the sides; adding to the sides takes away from the middle. The sweet spot is the maximum likelihood model. It is one in which we give more credence to things that happened more, and (because of the way probability works) give less credence to things that happened less.

This is a generalization of the principle of restrictiveness to the case where we have a gradient notion of “consistency with the data”: what restrictiveness says is “pick the model that predicts the observed data and as few other logically possible occurrences as you can”; what maximum likelihood says is “pick the model that gives the highest probability possible to the data, and (in so doing) the lowest probability possible to other logically possible occurrences.”

A relation between maximum likelihood and restrictiveness principles has been pointed out before (Collet, Galves & Lopes 1995; Jarosz 2006; Foraker, Regier, Khetarpal,

Perfors & Tenenbaum 2009). The value of the likelihood will be related to the “goodness” of fit (although it will not necessarily determine it) in virtually any probabilistic inference scheme, so that the mere use of probability to do inference will not only predict restrictiveness, but explain it (in the sense that it will come automatically).

Simplicity, on the other hand, is something rather different. It comes for free with Bayesian inference, but not necessarily with other ways of doing probabilistic inference. It is worth correcting one mistake in Tenenbaum 1999 relating to simplicity, restrictiveness, and the Bayesian Occam’s Razor, before we spend the rest of this section introducing the Bayesian Occam’s Razor alone. Tenenbaum introduces something called the size principle in the context of a simple Bayesian framework for concept learning, where “concepts” are just collections of integers. If the choice for the learner, unlike in the example above, is between different finite sets of integers, rather than different normal distributions, then the likelihood function will change, but the same principles of maximum likelihood will apply. In particular, if the likelihood function simply says expect any of the numbers that belong to the concept equally, then, multiplying the function through once for each of our observations, we get this composite likelihood for  $N$  data points observed under the assumption of a concept  $H$  with  $\text{Size}(H)$  integers in it:

$$(30) \quad \Pr[X|H] = \left[ \frac{1}{\text{Size}(H)} \right]^N$$

Just as with the normal distributions, it is better to pick concepts that do not predict unobserved elements to satisfy maximum likelihood. Thus, if there are different possible concepts that all predict the data should be at least possible, but some are larger than

others, then we should pick smaller concepts: here, restrictiveness aligns with simplicity. But this is a coincidence. It is not the same thing, because, in general, restrictiveness does not align with simplicity. Tenenbaum states that the size principle is the same as the Bayesian Occam's Razor discussed by MacKay 2003; but the Bayesian Occam's Razor is not due to the likelihood alone, and in fact, it can be thought of as a kind of general law about the behavior of prior distributions. If one were to add a prior of the sort that would give rise to a Bayesian Occam's Razor effect (we will elaborate on what that would look like later, but in this case an example would be most priors over sets of multiple concepts, not all containing the same number of concepts) then in fact both laws of inference would separately encourage smaller concepts in Tenenbaum's framework.

### 2.5.2 Model evaluation in statistics

The Bayesian Occam's Razor arises in hierarchical Bayesian inference, which is best understood by looking at a common hierarchical practice called model evaluation. Model evaluation, in turn, is best understood in contrast with parameter estimation. Parameter estimation is the most intuitive operation in statistical inference. It is what we are doing with any procedure that takes a collection of data as input and returns a model for that same data. Statistics students are almost always given this concept first, whether they realize it or not. The first exercise in almost every statistics class is estimating the location parameter of a normal distribution. This usually goes under the unfortunate label of "estimating a population mean"—a presentational move which is an attempt to go from a concept which is meaningless to students, "location of a normal distribution," to one that

is slightly more familiar because it includes the word “mean,” which students understand as “average.” This presentation is problematic because “finding the average” is a very misleading way for students to understand what it is they are doing in the example. To properly understand, students need to see the example from the point of view of scientific reasoning, not a mathematical operation. In what follows I explain parameter estimation using this example; then I contrast parameter estimation with model evaluation.

Tineke	175
Ineke	183
Anneke	178
Aaltje	192
Marietje	203
Catharijntje	180
Willemijntje	188
Leentje	183
Maaïke	192
Astrid	177

Table 2.1: Made-up heights of ten adult Dutch females in centimetres.

Suppose that the heights of ten adult Dutch females are collected: see Table 2.1. The student is told that one could easily assume that this kind of data “follows a normal distribution”; and that it turns out a good way to estimate this normal distribution’s mean (which is called the “population mean”) is just to compute the arithmetic mean of these ten numbers (185). For the student to properly understand this, the link between the ten numbers we are given and the function that yields the bell-shaped curve—called the normal density function, but let us simply call it  $f$ —first needs to be made crystal clear. What we are attempting to do in using a parametric distribution is to make statements that go beyond the data. The student is being told that there is a number  $\mu$  that can be filled in which will fix the function  $f$  at some particular point on the number line; among other



things, it is a good guess at where the maximum of the curve might be placed on the line.

Here is that function for concreteness:

$$(31) \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The function  $f$  is a function of a height  $x$  that can help us go beyond the ten measurements, to do two related things: calculate how many times more frequent one height should be than another (the ratio  $\frac{f(x_1)}{f(x_2)}$ ); and calculate how frequent a given range of heights should be (the area under the curve from  $x_{\perp}$  to  $x^{\top}$  gives a number that can be read off as a percentage). The idea behind the word “should” is that the normal distribution is giving us predictions, which we understand as being somehow hypothetical; or, at least, the normal distribution, including the answers we get out of it, if not hypothetical or theoretical, has some ontological status quite different from our ten data points, or in fact any heights we could actually observe.

The student has then been told that the average—which seems quite concrete, some sort of numerical summary of the observed heights—is a good substitute or estimate for  $\mu$ , although it is not really  $\mu$ . Students generally stop understanding around this point, because they have been told nothing of any substance about what the function  $f$  actually is supposed to mean; but assigning it some interpretation is crucial for understanding the difference between parameter estimation and model evaluation.

For example, one interpretation of  $f$  and its fixing-point, or location,  $\mu$  is as empirical quantities that cannot be known directly. Perhaps if we looked at the Dutch genome, we could do some calculations that would tell us what the exact adult height of a female

would be if the developmental program unfolded under some ideal circumstances; then we could list and model all the circumstances that influence development, and they would turn out to satisfy the requirements for the result following a normal distribution.<sup>10</sup> That would give us another number, which is called  $\sigma$  and tells us the exact shape of the bell curve, which also needs to be inserted before we can use the function  $f$ . Even though the numbers  $\mu$  and  $\sigma$  are derived quantities, they are still in this scenario ultimately real quantities about which we can in principle be right or wrong, because the Dutch genome is real and therefore has real propensities to behave in certain ways. This is one answer that we could give the student: the location of the normal distribution is of interest because it is some actual physical quantity.

This whole idea of interpreting the parameters objectively is itself problematic, however, if we do not believe in such a thing as “the Dutch population”—suppose there is no firm genetic boundary delimiting Dutchness even at a given instant in history. Then no distribution, normal or otherwise, can ever represent an objective law of (Dutch) nature, knowable or not, because there is no such objective thing as “Dutch nature,” which in this case means there nothing that Dutch women actually have in common that makes them Dutch. The question is why, in the face of this knowledge, one would continue to study the Dutch at all, and the answer is simply to take the idealization one step further: if one imagined that there were such a thing as “Dutch,” here is what it might be; for any given point in time, one can look at some conveniently defined finite population of Germanic-blooded people (say, the “Dotch” people) and construct a way of translating between this mathe-

---

<sup>10</sup>Put aside that the Central Limit Theorem is an asymptotic result, and the fact that the distribution is surely truncated well above zero; the point is that there would be some distribution that could be computed if we had all these facts about the world.

mathematical fantasy and the statistical reality of this population. It may be easier to model the Dutch than to pay close attention to the Dutch. The imaginary facts about the imaginary Dutch would then be a purely instrumental, as opposed to a realist, model of the actual Dutch: similar enough to something real that they are useful in approximating proportions and relative frequencies.

In either case, once we assign the normal distribution some interpretation, the point of the exercise of parameter estimation becomes quite intuitively clear. There is a quantity, somewhere in some corner of the actual world, or of the world of ideas, that is supposed to help us to understand something, to give us a model of something real; but we want do not know what that number is, so we must make a guess, informed by some observations.

The concept of model evaluation is very different. In an experiment, we may have two conditions in which we measure something like response time. The process of finding the means of our response times, and then deriving the difference between the two means, or then an “effect size” summarizing this difference in a standardized way, could be seen as parameter estimation, once we introduce the idea that the reaction times have some idealized distribution that can be interpreted as a model of what is going on in our experimental subjects’ behaviour, instrumental or realist. (Until we introduce such a model it is just a convenient numerical summary.) However, we may then wish to do a significance test, involving something called a  $p$ -value, to see if the two conditions are “really” different; this is not an act of parameter estimation, but something else. Or we may have a more complex experimental design, with multiple “factors,” each of which we have incorporated into our experimental materials in a particular way in order to test a more complicated hypothesis—say the factors are the frequency of a word, its length, and its morphological

status, simple or complex, and we have set up a response-time experiment to investigate if there is really any cognitive reality to morphological status. We start by “fitting a linear model” in which there are different numerical effects of different factors, and of their interaction. This is parameter estimation—and again, we can interpret the model as we see fit in order to understand what the resulting estimates represent. Then, however, we want to ask something about whether there is “really” an effect of morphological status; there is a number in the result of the model fit which represents this effect quantitatively, but it is not enough to look at it and see if it is exactly zero. We must do something else. There are also significance tests here, but, to illustrate the range of possibilities, there is also another way of reasoning about similar issues, which is something that often goes under the heading of model comparison—fit the model with and without the formal term for morphological status, and then compute a comparison statistic and see if it crosses some threshold (one familiar to experimenters will be the Bayesian Information Criterion, or BIC, ratio). Although significance testing and model comparison have results that need to be interpreted in very different ways, we can still get the idea that these are sorts of meta-level procedures, which we collectively call model evaluation tools.

The general idea is that our parameter estimation procedure only gives reasonable results contingent on some set of modelling assumptions, but part of what we are interested in is actually those assumptions themselves. One strange consequence of this is that there will often be two different ways of asking what seems like the same question: are the two conditions in our experiment the same or not? We could see if the parameter estimates come out exactly the same in the two different conditions, or we could compute a  $p$ -value and see if it falls below a standard threshold like 0.05; we would in general get

two different answers. With the right concrete interpretation of our parameter estimation, this can actually make perfect sense: if we are looking at Dutch and German people, we might want to assume they are two different groups and infer something about the “ideal height” implied by their respective genomes (which could still be accidentally identical), or we might want to infer something about whether they are actually two different groups genetically, which then allows for two “discrepant” parameter estimates—the genomes are different as regards the implied ideal height, but if we were to simply assume they were different, we might find the best parameter estimates come out the same; or, they are not different, although our limited data would lead us to believe otherwise (this is the case we usually try to rule out in statistical tests). The gap between the parameter estimates and the true values opens up the possibility of either type of discrepancy; and the possibility that the two groups can be different in some way that we do not have information about and still be accidentally identical with respect to the property the parameter represents opens up another avenue to the first type. Thus it makes intuitive sense that we would want to find some way of evaluating different sets of modelling assumptions statistically. This is how parameter estimation differs from model evaluation.

Some terminology: the objects of inference in parameter estimation are parameter values or model instantiations; the objects of inference in model evaluation are competing model frameworks.

### 2.5.3 Bayesian inference and model evaluation

Bayesian inference uses the posterior distribution over parameter values to assess the “value” of those parameter values in the sense discussed earlier. We can talk about this in terms of evaluation measures, and, as discussed above, this lets us derive a relation between the prior evaluation measure and the posterior evaluation measure once we are given the likelihood function, using Bayes’ Rule. Remembering that in the context of learning the likelihood function takes grammar–data set pairs and gives back a “degree of consistency,” we can make this concrete, looking at the example with the Dutch women in this light—swapping out “grammar” for “model,” and moving temporarily from the realm of cognition to scientific inference. We would say that the normal density function, seen as a function now of both  $x$  and  $\mu$ —imagine that  $\sigma$  is known in advance—is the likelihood function; or, if we knew ten heights, as we did above, we could construct a function of  $\mu$  values paired with sequences of observations,  $x_1, \dots, x_{10}$ , by simply multiplying together the normal density function 10 times, once for each height  $x_i$ . In either case, the likelihood function would be telling us how well some guess at the value of  $\mu$ , whatever we need it to represent for our purposes, would “fit,” or be consistent with, the observations. To do Bayesian inference, we then need to pick a prior measure—in this case some probability distribution that takes as input  $\mu$  only. In the scientific context this is harder to understand than in the cognition context, but, in either case, it will represent a bias in the general sense. To do parameter estimation using Bayesian inference, one simply decides on a way of using the posterior that is derived from of the likelihood plus the prior to pick the “best” value of  $\mu$ —very often the one with the highest posterior probability, but sometimes other

things, like the mean value of  $\mu$  according to the posterior.

Model evaluation, under the Bayesian approach, is simply an extension of parameter estimation. The Bayesian approach to model evaluation is to add an additional parameter—call it  $\omega$ —to code for different model frameworks. To distinguish  $\omega$  from more “basic” parameters like  $\mu$ , it is often called a hyperparameter; however, the same Bayesian methods are used to compare values of  $\omega$  as would be used for any other parameters.

To take an example, consider now a common type of statistical model, alluded to briefly above: the linear model. We discussed above the hypothetical example of an experiment in which we manipulate word length, frequency, and morphological complexity, and measure response time as a dependent variable. A common and very simple type of parameter estimation for this data—first without the hyperparameter—would be one that operates under the assumption of this model:

$$(32) \quad y - \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Here,  $y$  is the dependent variable, and  $x_1$ ,  $x_2$ , and  $x_3$  are the three independent variables; they might be binary, representing different conditions, (coded in some carefully chosen but arbitrary way, like 0–1), or they might be real-valued—for our purposes it does not matter. The values  $\beta_0$  and  $\beta_1$ – $\beta_3$  are the parameters to be estimated, which in this case are called the regression coefficients. The idea in a linear model is that if we could subtract some random error,  $\varepsilon$ , from each observation, we would see the perfect linear relation between independent and (corrected) dependent variables given by the equation. The error  $\varepsilon$  follows a normal distribution with location zero in the standard linear model,

which is mathematically convenient and gives certain standard procedures for parameter estimation interesting alternate interpretations (in particular, the “maximum likelihood” procedure also has the “minimize the distance to the line” interpretation). Bayesian approaches compute a posterior distribution over the set of possible parameters, in this case the set of real-valued quadruplets  $\langle \beta_0, \beta_1, \beta_2, \beta_3 \rangle$ .

To keep things very simple, suppose that we can get away with only considering the effects of word length and morphological complexity, and that each of the influences on the response time can either be exactly numerically equal to the independent variable, or else have no effect at all. In other words, consider this simpler model in which we have removed the term  $\beta_0$  (the intercept) and one of the coefficients (the one for frequency, by whatever arbitrary convention we use to label the independent variables):

$$(33) \quad y - \varepsilon = \beta_1 x_1 + \beta_2 x_2$$

By what we have said we have then restricted the parameters (a pair, or two-dimensional parameter vector,  $\langle \beta_1, \beta_2 \rangle$ ) to the set  $\Theta = \{ \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 0, 1 \rangle, \langle 1, 1 \rangle \}$ . The most intuitive way of computing the posterior is to move through all four possibilities. The likelihood would be computed, for each possibility, by filling in the values  $y$ ,  $x_1$ , and  $x_2$  associated with a particular observation, rearranging and subtracting to obtain  $\varepsilon$ , then applying the normal density function. (Again we are ignoring the parameter  $\sigma$  that we would need to actually do this evaluation, but in principle we could evaluate this in a Bayesian way, too, by moving to a set of three-dimensional parameter vectors, with a third for  $\sigma$ .) To extend this to multiple data points, as alluded to before, we would simply multiply all the values



together for a given hypothetical parameter vector  $\langle \beta_1, \beta_2 \rangle$ . The validity of this move requires the assumption that the data points are “independent” of each other in a technical sense, and only that assumption; again, this will be elaborated on further below.

To derive the posterior, we then multiply in the prior. Letting  $L(X, \beta_1, \beta_2)$  be the likelihood as just explained, the posterior for, say,  $\langle 1, 1 \rangle$  is:

$$(34) \quad \frac{L(X, 1, 1) \cdot \Pr[\langle 1, 1 \rangle]}{\sum_{\langle \beta_1, \beta_2 \rangle \in \Theta} L(X, \beta_1, \beta_2) \cdot \Pr[\langle \beta_1, \beta_2 \rangle]}$$

If there were only one observation, say, an observation of 500 milliseconds ( $y = 500$ ) as a response time, for a word of length 5 ( $x_1 = 5$ ) in the morphologically complex word condition ( $x_2 = 1$ ), then we would get  $\varepsilon = 500 - 5 - 1 = 494$ . The likelihood would be computed using the normal density function, applying it to the number 494. (In addition to the fact that we need to know  $\sigma$  to do this, there is also the fact that response times to words—500 ms is a reasonable response time in the “lexical decision” task—do not appear to be on the same scale as either of our two dependent variables, which one would think ought not to make any difference to whether a model can be made to fit the data; it does in this toy example, and that is why we never actually use this hypothesis space where effects can only be exactly one and there is no intercept term.) In the denominator, we need to sum over all values of this quantity for all possible values of  $\langle \beta_1, \beta_2 \rangle$ . Suppose the prior were uniform:  $\frac{1}{4}$  for each of the four parameter values, which means no a priori bias to any of them. Then we would see that the prior term cancels, and we get the following,

where  $\phi$  is the normal density function:

$$(35) \quad \frac{\phi(494)}{\phi(500) + \phi(499) + \phi(495) + \phi(494)}$$

The posterior probability of  $\langle 1, 1 \rangle$  would be this fraction, and for any other parameter value we can do the same. For the sake of having some numbers, we can plug in  $\sigma = 100$  and compute, and we find that we have a change in the subjective probabilities of some ideal observer—from a uniform distribution with  $\Pr[\cdot] = 0.25$  across the board—to the distribution  $\langle 0, 0 \rangle : 0.21, \langle 0, 1 \rangle : 0.22, \langle 1, 0 \rangle : 0.27, \langle 1, 1 \rangle : 0.29$ . This updating is the basic operation we use to do parameter estimation using Bayesian inference.

We can then introduce our hyperparameter  $\omega$ . Remember that model evaluation compares different modelling assumptions. Here we can let  $\omega = 1$  stand for the model we just examined; in principle  $\omega = 2$  could be anything at all, but suppose it stood for this interesting model:

$$(36) \quad y - \varepsilon = \beta_1 x_1$$

How do we connect this with Bayesian inference? The idea is to see the selection of models as a restriction on the parameter values. In this case, the restriction is mathematically equivalent to setting  $\beta_2$  to be necessarily zero. For the probabilistic agent, what this means is that, if  $\omega = 2$ ,  $\Pr[\langle 0, 0 \rangle] = \Pr[\langle 1, 0 \rangle] = 0$ , but if  $\omega = 1$ , this is not the case, and there is some non-zero probability assigned to the parameter vectors with  $\beta_2 = 0$ : the prior on  $\langle \beta_1, \beta_2 \rangle$  changes depending on the value of  $\omega$ , and changing the prior can have

the same effect as making the inference restricted to a particular subset of the hypothesis space, by making use of the minimum value for probabilities, 0. No matter what our model comparison, we can always find a way of contriving to make our model comparison a choice of different priors in this way.

The formal apparatus is as follows: first, the question we are asking is now different. The posterior of interest is  $\Pr[\omega = 1|X]$ ,  $\Pr[\omega = 2|X]$ . The prior distribution on  $\langle\beta_1, \beta_2\rangle$  we were discussing before is not the prior anymore; it is a conditional probability distribution, which will be of use to us, but in a different way. Although we used the uniform distribution as an example, it could have been anything—the point is that this distribution is now treated as  $\Pr[\langle\beta_1, \beta_2\rangle | \omega = 1]$ , the “conditional probability of  $\langle\beta_1, \beta_2\rangle$ , given  $\omega = 1$ .” The distribution  $\Pr[\langle\beta_1, \beta_2\rangle | \omega = 2]$  could also be anything, although in our case it needs to be some distribution with  $\Pr[\langle 0, 0\rangle | \omega = 2] = \Pr[\langle 1, 0\rangle | \omega = 2] = 0$ . This can be used to compute the posterior.

Using Bayes Rule to compute the posterior for  $\omega = 1$  requires that we find a probability distribution proportional to the following:

$$(37) \quad \Pr[X|\omega = 1] \Pr[\omega = 1]$$

The law of total probability (which actually follows from the basic axioms of probability theory) allows us to expand this in terms of  $\langle\beta_1, \beta_2\rangle$ . It says (in one simple form): given some conditional probability distribution  $\Pr[A|B]$ , partition the whole space of events  $B$  that we can condition on in some way, say  $\{B_1, \dots, B_K\}$ .  $\Pr[A] = \sum_{i=1}^K \Pr[A|B_i] \Pr[B_i]$ .

This gives us a way to expand  $\Pr[X|\omega = 1]$ :

$$(38) \quad \Pr[X|\omega = 1] = \sum_{\langle \beta_1, \beta_2 \rangle \in \Theta} \Pr[X|\langle \beta_1, \beta_2 \rangle \ \& \ \omega = 1] \cdot \Pr[\langle \beta_1, \beta_2 \rangle |\omega = 1]$$

Notice that the distributions we employ remain conditional on  $\omega = 1$ ; this is just because the law of total probability applies equally to conditional distributions as to any other. The probability of the data (the likelihood) is mediated only by the model instantiation, however, and not the model framework. Thus we can drop that part of the condition, and for the whole expression, we obtain:

$$(39) \quad \Pr[X|\omega = 1] = \sum_{\langle \beta_1, \beta_2 \rangle \in \Theta} \Pr[X|\langle \beta_1, \beta_2 \rangle] \cdot \Pr[\langle \beta_1, \beta_2 \rangle |\omega = 1]$$

What acts as the likelihood is actually an average of the likelihood values for each of the possible parameter vectors, weighted not equally but by the conditional prior probability of each of those parameter vectors. Because we are governed by the law of total probability, we have no choice but to derive the likelihood for  $\omega = 1$  from the individual likelihoods in this way (which also happens to be quite convenient). To see this again, let us change to a more readable notation: we will write a subscripted  $\lambda$  for the likelihood given some data set, where the subscript tells us what the parameter vector we are assuming is—so that  $\lambda_{0,0}$  is the likelihood for some data given the model  $\langle \beta_1, \beta_2 \rangle = \langle 0, 0 \rangle$ , for example; and for the conditional prior on  $\langle \beta_1, \beta_2 \rangle$  we will write  $p$ , similarly subscripted, and with a superscript to indicate the model framework, so that  $p_{0,0}^{(1)} = \Pr[\langle 0, 0 \rangle |\omega = 1]$ .

What we are saying is that the term  $\Pr[X|\omega = 1]$  expands as follows:

$$(40) \quad \lambda_{0,0}p_{0,0}^{(1)} + \lambda_{1,0}p_{1,0}^{(1)} + \lambda_{0,1}p_{0,1}^{(1)} + \lambda_{1,1}p_{1,1}^{(1)}$$

This is what makes model evaluation possible in a Bayesian setting: we add a layer to the inference, another variable upon whose value the model instantiation is contingent—not determined by it, but influenced by it, via a change in its prior distribution. What we have is a hierarchical model, with the model framework  $\omega$  occupying a higher “level” than the model instantiation  $\langle \beta_1, \beta_2 \rangle$ . We will finish this section by using this setup as a concrete example of the law of inference we want to derive: prefer simpler model framework—the Bayesian Occam’s Razor.

Consider comparing the two model frameworks by taking the ratio of the two posterior values,  $\Pr[\omega = 1|X]$ ,  $\Pr[\omega = 2|X]$ . We could easily check to see how many times more a posteriori probable one framework was than another in this way. The prior  $\Pr[\omega = 1]$ ,  $\Pr[\omega = 2]$  is some bias over model frameworks, but suppose the prior on  $\omega$  is uniform. Then it will cancel, and we will be left with this ratio:

$$(41) \quad \frac{\lambda_{0,0}p_{0,0}^{(2)} + \lambda_{1,0}p_{1,0}^{(2)} + \lambda_{0,1}p_{0,1}^{(2)} + \lambda_{1,1}p_{1,1}^{(2)}}{\lambda_{0,0}p_{0,0}^{(1)} + \lambda_{1,0}p_{1,0}^{(1)} + \lambda_{0,1}p_{0,1}^{(1)} + \lambda_{1,1}p_{1,1}^{(1)}}$$

Now, what we know about our restriction in model  $\omega = 2$  is that it has the effect of removing two terms:

$$(42) \quad \frac{\lambda_{0,0}p_{0,0}^{(2)} + \lambda_{1,0}p_{1,0}^{(2)}}{\lambda_{0,0}p_{0,0}^{(1)} + \lambda_{1,0}p_{1,0}^{(1)} + \lambda_{0,1}p_{0,1}^{(1)} + \lambda_{1,1}p_{1,1}^{(1)}}$$

Here is the key intuition which we will carry throughout the chapter: probability distributions, including each of the two individual conditional prior distributions here, must sum to one, and so the prior probabilities in the numerator are “compressed” with respect to those on the bottom. That is,  $p_{0,0}^{(1)} + p_{1,0}^{(1)} = 1$ , but  $p_{0,0}^{(2)} + p_{1,0}^{(2)} \leq 1$ . The idea is to leverage this fact in such a way as to make it guaranteed that the model framework in the numerator—the simpler model, the one with the more restricted subset of the hypothesis space—takes the same likelihood values,  $\lambda_{0,0}$  and  $\lambda_{1,0}$ , and weights them by larger numbers than the more complex model framework in the denominator. Then we can rightly say that the simpler model will be necessarily preferred—will have higher posterior probability—all things being equal (what exactly this means will be spelled out momentarily).

It turns out that all we need to do is to assume that the change of model framework leaves the relative preferences for different parameter values intact. That is to say, over some shared subset of  $\Theta = \{\langle 0,0 \rangle, \langle 1,0 \rangle, \langle 0,1 \rangle, \langle 1,1 \rangle\}$ , the model frameworks at least agree on how many times more a priori likeli one parameter vector is than another. In this case, we want the two model frameworks to give the same ratio of conditional prior probabilities for the two “restricted” values, so  $p_{0,0}^{(2)} / p_{1,0}^{(2)} = p_{0,0}^{(1)} / p_{1,0}^{(1)}$ ; and this, by a bit of algebraic manipulation, is equivalent to saying that  $p_{0,0}^{(2)}$  and  $p_{1,0}^{(2)}$  are equal to the corresponding values  $p_{0,0}^{(1)}$  and  $p_{1,0}^{(1)}$ , multiplied by  $\frac{1}{p_{0,0}^{(1)} + p_{1,0}^{(1)}}$ . The ratio becomes:

$$(43) \quad \frac{1}{p_{0,0}^{(1)} + p_{1,0}^{(1)}} \cdot \frac{\lambda_{0,0}p_{0,0}^{(1)} + \lambda_{1,0}p_{1,0}^{(1)}}{\lambda_{0,0}p_{0,0}^{(1)} + \lambda_{1,0}p_{1,0}^{(1)} + \lambda_{0,1}p_{0,1}^{(1)} + \lambda_{1,1}p_{1,1}^{(1)}}$$

The two conditional prior terms in the numerator are identical to the first two terms

of the denominator, up to a scaling factor; the scaling factor is the ratio at left, and, notably, it is necessarily no less than one (because the bottom of the ratio is no more than one). This means that this scaling factor increases the preference for the reduced model; this is the essence of the Bayesian Occam's Razor. Prior probability distributions, like any probability distributions, must work with the same, finite amount of "probability mass"; less restricted distributions must spread the probability mass around, while for more restricted distributions, the mass is more concentrated. The scaling factor can be seen as the degree to which the probability is concentrated, and because of the laws of probability theory, the scaling factor must always act to increase the a priori preference for the simpler model.

It is not guaranteed that the simpler model will be preferred; the scaling factor is contributed only by the model instantiations shared by the two model frameworks, and so if those parameter vectors do not fit the data very well, then the scaling factor will need to compete with the strong preference for the more complex model brought on by the right-hand ratio. The less the additional parameters made available under the complex model contribute to the value in the denominator, the more the Bayesian Occam's Razor effect will be seen. This contribution will be small if the additional parameters are a priori considered highly implausible, or they yield poor-fitting model instantiations compared to the narrower set of parameter values. Whether the reduced model will be preferred depends on precisely how large the scaling factor is, and precisely how much the additional parameters contribute to the evaluation. This is the powerful sense in which, all things being equal, Bayesian model evaluation will prefer the simpler model.

Here is what we have done: we have constructed an example of Bayesian model evaluation, and we have shown that this construction leads to a bias—a preference in the

prior distribution, broadly construed—for simpler models. There is an intuition that this construction is no accident; what got us the result was simply that we found a way to leverage the fact that, for the simpler model, the prior probability is more highly concentrated on certain values. Although we have not proven it yet, this Bayesian Occam’s Razor is general enough to be considered a law of inference. In the next section, we spell out the details of the law, including the conditions under which it holds.

## 2.5.4 Conditions for a Bayesian Occam’s Razor

We now give a strong condition under which a BOR will emerge which substantially generalizes the individual examples of the BOR presented in Jaynes 2003, MacKay 2003. Although it is possible to see analogous effects under more general circumstances, the condition represents a pair of model frameworks which alter the prior bias as little as possible, a property which can be seen as an optimality condition on the evaluation measure.

- (44) Framework consistency principle (FCP): For any pair of model framework parameter values  $\omega_1, \omega_2$ , the prior distributions to which they correspond must be similar over some subset of their respective parameter spaces.
- (45) Framework consistency principle + data (FCPD): For any pair of model framework parameter values  $\omega_1, \omega_2$ , the prior distributions to which they correspond must be similar over some subset of their respective parameter spaces with respect to the likelihood function.



Intuitively: two model frameworks may define radically different types of models; however, they satisfy the FCP if there is some subset of these model instantiations which are similar across the two frameworks. “Similar” will mean that the relative prior probabilities of all possible parameter values, or sets of parameter values, remain the same across the two models—that is, within the similar set, the biases for one model instantiation over another are the same. “Similar with respect to a likelihood function” will add the additional clause that this subset of the models is the same across the two models with respect to how each of the model instantiations treats the given set of data.

In a simple case, such as the one we considered in the previous section, the two sets of parameters will be identical across the two model frameworks. In this case, a very simple version of similarity is:

- (46) Similarity (preliminary): Two probability distributions  $\Pr_1$  and  $\Pr_2$  over the common parameter space  $\Theta$  are similar over some  $A \subseteq \Theta$  if, for all  $S \subseteq A$ ,
- $$\Pr_1 [S|A] = \Pr_2 [S|A].$$

Assuming that the likelihood is dependent only on  $\bar{\theta}$ , and not  $\omega$ , (capturing the fundamental intuition behind model evaluation), this is sufficient to satisfy both descriptions above. In particular, this says that the conditional probability given membership in the similar set  $A$  is the same across the two distributions, where this is defined as  $\Pr_1 [S] / \Pr_1 [A]$  (similarly for  $\Pr_2$ ). This is the only way to satisfy the condition that the relative probabilities remain the same.<sup>11</sup>

---

<sup>11</sup>The only reasonable relative-difference condition is that  $\Pr_1 [S] / \Pr_1 [T] = \Pr_2 [S] / \Pr_2 [T]$ , which includes the case where  $T = A$ .

Generalizing this to the case where the two parameter spaces,  $\Theta_1$  and  $\Theta_2$  are distinct, we obtain:

(47) Similarity: Two prior distributions  $\Pr_1$  and  $\Pr_2$  over  $\Theta_1, \Theta_2$  are similar for some  $A_1 \subseteq \Theta_1, A_2 \subseteq \Theta_2$  if there is a bijection  $f : A_1 \rightarrow A_2$  such that, for all  $S \subseteq A_1$ ,  $\Pr_1 [S|A_1] = \Pr_2 [f(S)|A_2]$ .

(48) Similarity with respect to a likelihood: Two prior distributions  $\Pr_1$  and  $\Pr_2$  over  $\Theta_1, \Theta_2$  are similar for some  $A_1 \subseteq \Theta_1, A_2 \subseteq \Theta_2$  with respect to a fixed likelihood  $\lambda(X|\cdot)$ , if there is a bijection  $f : A_1 \rightarrow A_2$  such that, for all  $S \subseteq A_1$ ,  $\Pr_1 [S|A_1] = \Pr_2 [f(S)|A_2]$ , and  $\lambda(X|\theta) = \lambda(X|f(\theta))$ , for all  $\theta \in A_1$ .

Since  $\Pr_1 [A_1|A_1] = \Pr_2 [A_2|A_2] = 1$ , it is for all practical purposes assured that  $f(A_1) = A_2$  (where  $f$  applied to a set means the image of the set under  $f$ ). For any pair of model framework parameter values  $\omega_1, \omega_2$ ,  $\Pr [X|\omega_1] = (\Pr [A_1|\omega_1] / \Pr [A_2|\omega_2]) \cdot \Pr [X|\omega_2]$ . Applying this to the example of the previous section, we find we need not appeal to the more complex version, however; the scaling result holds immediately.

Assuming that FCPD holds, then, more generally, for nested models ( $A_1 = \Theta_1, A_2 \subseteq \Theta_2$ ), we have the following Bayes factor (applying the same notation from the previous section):

$$(49) \quad \frac{1}{\Pr [A_2|\omega = 2]} \cdot \frac{\int_{A_2} \lambda_{[\cdot]} dp_{[\cdot]}^{(2)}}{\int_{A_2} \lambda_{[\cdot]} dp_{[\cdot]}^{(2)} + \int_{\bar{A}_2} \lambda_{[\cdot]} dp_{[\cdot]}^{(2)}}$$

As before, the right-hand ratio is the relative weighted fit of the reduced model as compared to the full model (a function of the fits of the individual parameter vectors and

their prior probabilities); it can never favor the reduced model and must be at most one. The left-hand ratio is the relative prior probability of the parameter vectors permitted under the reduced model, taken as a set. Since this is by definition one under the reduced model, this ratio can never favor the full model: it must be at least one.

In the most general case, where the model frameworks overlap for some similar set, but neither is nested within the other:

$$(50) \quad \frac{\Pr[A_1|\omega=1]}{\Pr[A_2|\omega=2]} \cdot \frac{\int_{A_2} \lambda_{[\cdot]} dp_{[\cdot]}^{(2)} + \frac{\Pr[A_2|\omega=2]}{\Pr[A_1|\omega=1]} \int_{A_1} \lambda_{[\cdot]} dp_{[\cdot]}^{(1)}}{\int_{A_2} \lambda_{[\cdot]} dp_{[\cdot]}^{(2)} + \int_{A_2} \lambda_{[\cdot]} dp_{[\cdot]}^{(2)}}$$

In this case, it is not obvious which model is “larger,” where larger now means having more prior probability mass assigned to the complement of the similar set; the denominator model is simply the one in terms of which the ratio is written, and has no special status. The numerator now contains an additional term, boosting the weighted fit to account for the parameters which do not correspond to any in the denominator model; meanwhile, the scaling factor is reduced by multiplying in the prior weight of the similar set under  $\omega = 1$ , now no longer necessarily one.

In sum, any prior distribution over model frameworks and parameter values which has the FCPD property will show the BOR effect, now generalized to (50). Although it is incoherent to say that the BOR effect shows up without the need to incorporate a bias for either model (since the BOR is a part of the prior, and thus is itself a bias), it is reasonable to say that the bias arises in a non-arbitrary way: the FCP property will hold whenever the prior is assigned according to the general principle that the bias toward different, equivalent model instantiations should not change as a function of the model

framework, and the FCPD property will hold whenever this happens to have no effect on the predictions for a particular data set.

Note also that none of this tells us which model is “larger” on the basis of anything apart from the prior measure itself; this is actually rather difficult when two model frameworks both give rise to infinite conditional hypothesis spaces of the same cardinality. We will return to discuss this issue further when we apply the BOR to grammar frameworks, which means much of the next section will be dedicated to being specific enough about what it means to specify a grammar that we can associate the “subset” of grammars with an independent notion of “lower complexity”; however, for the time being, we can simply attend to the crucial idea: one model framework must be “more complex” than another when its prior assigns more probability mass outside a set which is shared (by a bijection) between the two. This leads to the Bayesian Occam’s Razor for the same reason that maximum likelihood principle gives rise to restrictiveness-like effects.

## 2.6 The Optimal Measure Principle

The goal of this analysis, as it continues, is to highlight the circumstances under which Bayesian inference for grammars will give rise to a simplicity bias. We now know a good way to identify cases of BOR: look for hyperparameters which imply a change in the “size” (in terms of the uncertainty in the prior measure) of the set of possibilities for other parameters in the specification of a grammar; the clear cases will be those for which these other parameters’ distributions are independent of the choice of hyperparameter. The point is deeper, though, as we wish to show that such priors are “natural” given only the

statement of “what a grammar looks like.” In Chapter 1, I raised the question of what it means for a grammar to “look like” something anyway, raising the example of OT versus SPE type grammars, both of which can be “compiled out” to finite-state transducers—so aren’t they equivalent? Now, in that case, the three different versions of any given phonological mapping will also yield some structure, the “trace” of the computation, which will not be isomorphic across the three intensions. But the CG versus CFG case does not have this property. The question is, is the way we write out the grammar per se meaningful? The traditional answer in generative grammar was yes, (Chomsky 1957, Chomsky 1965, Chomsky & Halle 1968). Now, it need not be the case that the notation we choose is something we interpret as meaningful. But it does need to be the case, if grammars are non-atomic, that the relevant cognitive systems necessarily implicitly assert something about how to interpret the pieces of a grammar, implying an assertion about what those pieces are (e.g. in P&P, each grammar is a complex of parameter values, and in Aspects grammars, each grammar is a complex of PS and transformational rules)—and thereby implying an assertion of some “structure” over the set of grammars. That is what the notation should ideally capture, and, once we see this clearly, it will be obvious why the BOR should hold for hierarchically structured sets of grammars.

### 2.6.1 Formalizing grammars preliminary: Transparency and structure

In this section I introduce the idea that, whenever we put forward a theory of how language works cognitively, we are implicitly saying something about a structure over the set of grammars. Before saying anything about what a structure is, what a grammar is, or

anything else from first principles, let's consider an illustrative example.

In rule-based phonology, it is assumed that a grammar consists of a collection of rules each of the following form:

$$(51) \quad A \rightarrow B/C-D$$

To review: the phonological representation of a form is a sequence of symbols, and a rule of this form (roughly) replaces a subsequence matching *CAD* with *CBD*. For example, the rule  $i \rightarrow e / -k$ ; as the symbols for segments such as *i*, *e*, and *k* are actually complex objects made up of collections of features in this theory, the formalism also allows for only particular features to be matched and changed (for example, a single rule might change *i* and *u* to *e* and *o* respectively by setting  $A = +\text{high}$  and  $B = -\text{high}$ ). With a fully explicit representation of the input and the grammar, this basically reduces to another case of the same thing, with some minor added complexity due to the fact that the collection of features in each phoneme must be sequentialized. The rules compose, and the order in which they compose is a part of the grammar too. There are independently motivated constraints on the ways in which rules can apply to their own output which have the salutary effect of preventing any rule system from being super-regular (Kaplan & Kay 1994). The set of all such grammars (“SPE grammars”) is the set of all sequences of such rules. This set is (enumerably) infinite because both *A*, *B*, *C*, and *D*, on the one hand, and the sequence of rules itself, have unbounded length.

Berwick & Weinberg (1984) introduce a notion of “transparency” for understanding how theoretical descriptions of grammars relate to actual performance systems. More

generally, we may consider how any two systems are or are not related (whether one is theoretical and the other is a real performance system, or both are actual cognitive systems) by considering what structure they share. A “structure” is just some collection of mappings which apply to elements of a set. Here are some different structures that could be constructed over the set of SPE grammars (not necessarily mutually consistent in any sensible way):

- (52) The grammar  $\langle i \rightarrow e / -k \rangle$  is a subpart of the grammar  $\langle i \rightarrow e / -k, \rightarrow k / -t \rangle$ , and so on.

[A single two-place function,  $\text{Subpart}(x, y)$ , mapping to  $\top$  or  $\perp$ ]

- (53) The grammar  $\langle i \rightarrow e / -k \rangle$  is a subpart of the grammar  $\langle i \rightarrow e / t - k \rangle$ , and so on.

[A single two-place function,  $\text{Subpart}(x, y)$ , possibly different from the previous]

- (54) The grammar  $\langle i \rightarrow e / -k \rangle$  is shorter than the grammar  $\langle i \rightarrow e / t - k \rangle$  which is shorter than the grammar  $\langle i \rightarrow e / -k, \rightarrow k / -t \rangle$ , and so on.

[A single two-place function,  $\text{Shorter}(x, y)$ ; or perhaps a one-place function,  $\text{length}(x)$ , mapping to a nonnegative integer or a member of some other set with an existing order; or, redundantly, both]

- (55) The grammars  $\langle i \rightarrow e / -k \rangle$ ,  $\langle e \rightarrow i / -k \rangle$ , and  $\langle i \rightarrow e / t - \rangle$  are, with some others, members of a set  $A_3$ ; the grammars  $\langle i \rightarrow e / t - k \rangle$ ,  $\langle e \rightarrow i / t - k \rangle$ , and  $\langle i \rightarrow e / -tk \rangle$  are, with some others, members of a set  $A_4$ ; and so on.

[A single one-place function,  $\text{Container}(x)$ , mapping either to a single set or to a set of container sets]

(56) The grammar  $\langle i \rightarrow e / -k \rangle$  has the extension  $\{\langle a, a \rangle, \langle ik, ek \rangle, \langle ds, ds \rangle, \dots\}$ , and so on.

[A single one-place function,  $\text{Extension}(x)$ , mapping to a set of input–output pairs]

We can see a linguistic theory as being correct if the set of grammars shares some, but presumably not all, structure with the human linguistic system. For example, a correct linguistic theory will almost certainly not capture cellular level details of how the brain works, but it will at least be a set of grammars with the correct extensions. Similarly, what we mean by “correct” in “correct extensions” is also “sharing some structure with the brain”—in that case with what we conventionally think of as static states representing inputs and outputs, as opposed to computations (possible grammars).

We like to think that some meaningful structure is captured in the notation we use for grammars, too: “choice of notations and other conventions is not an arbitrary or ‘merely technical’ matter ... . It is, rather, a matter that has immediate and quite drastic empirical consequences” (Chomsky 1965, 45). This obviously does not mean that there is a wax tablet or a paper tape inside the brain on which is written  $S \rightarrow NP VP$  (or whatever). It means that there is something inside the brain which shares some structure with  $S \rightarrow NP VP$ , structure which in the linguist’s grammar might be considered “notational.” In general we may think of various different sorts of higher-order properties of the string that codes the grammar as potentially meaningful in this way: not the appearance of the symbols themselves, but their combination, order, number, and so on; and, of course, the grammar will perhaps embed, and definitely imply, certain kinds of input and output representations, which for the linguist are sequences of symbols, but must, as we have



already said, share some structure with the brain's representations.

So what does it mean to “capture” or “share” structure? Sharing implies two sharers — so, start with two sets,  $S_0$  and  $T_0$ . Again, the structure associated with a grammar could be anything in this very general case where we are just trying to define “what is common” — length, ordering, a grouping into subsets, a set of pairs. Suppose the set  $S_0$  (in our case grammars) has an associated set of mappings  $f_1, \dots, f_n$ , each potentially mapping to some set other than  $S_0$  ( $S_1, \dots, S_m$ , for some  $m \leq n$ ; one here might be nonnegative integers, for example, representing length). These mappings constitute the “structure.” What we mean by “share” is that  $T_0$  also has  $n$  mappings associated with it, and they each “correspond” in some sense to a function of the structure of  $T_0$ . We will say that they correspond, or are shared, by virtue of a (generalized) homomorphism  $F$  which serves to map from  $S_0, S_1, \dots, S_m$  and  $f_1, \dots, f_n$ , on the one hand, to some other sets and mappings, on the other. One of these image sets must wind up being  $T_0$ , of course, and the mappings must “work the same” as the structure  $f_1, \dots, f_n$ . In particular:

$$(57) \quad a = b \Leftrightarrow F(a) = F(b)$$

$$(58) \quad f_i(a) = b \Leftrightarrow F(f_i)(F(a)) = F(b)$$

Some details:  $F$  is understood to be “overloaded” to apply, in the first case, to elements of  $S_0, S_1, \dots, S_m$  (“atoms”), and in the second case, to mappings as well. In words, (i)  $F$  preserves the distinctness of atoms (crucially, it does not collapse two atoms into one), and (ii)  $F$  preserves the action of all the mappings  $f_i$ . It is understood in the statement of the second axiom that  $F(f_i)$  applies to  $F(a)$ , which means that the domain  $Dom$

of  $F(f_i)$  is always at least the image of  $F$  as applied to  $\text{Dom}(f_i)$ , which we write as  $F(A)$  if  $f_i : A \rightarrow B$ ; the axiom implies then that the codomain of  $F(f_i)$  is at least  $F(B)$  (and what happens outside  $F(A)$  and  $F(B)$  is irrelevant so long as the action has been preserved for  $A$  and  $B$ ). So, if  $f_i : S_0 \rightarrow S_j$  then  $F(f_i) : F(S_0) \rightarrow F(S_j)$ ; by the first axiom,  $S_0$  and  $S_j$  remain distinct to the extent that they were distinct in the first place. That the action of the mappings is preserved is exactly what is meant by “structure is shared.”

Take a somewhat contrived example and we will see how this works; we will also see that in reality we go beyond demanding “some structure be shared” to saying exactly what structure and how that structure is shared. Categorical grammars and context-free grammars are two different formal devices for specifying sets of strings, and they are “weakly equivalent.” This means that, given any context-free grammar specifying a particular set of strings (say, for example, the grammar  $\{S \rightarrow aSb, S \rightarrow ab\}$ , specifying the language  $a^n b^n = \{ab, aabb, aaabbb, \dots\}$ ) there is a categorical grammar specifying the same set (in this case, one is  $\{a = A, S/B; b = S \setminus A, B \setminus S\}$ ), and conversely. Here are the pieces: for each type of grammar, we have a function which takes a given grammar to its string language, say  $L_{\text{CFG}} : G_{\text{CFG}} \rightarrow \mathcal{L}$ , from context-free grammars to sets of strings, and  $L_{\text{CG}} : G_{\text{CG}} \rightarrow \mathcal{L}$ , from categorical grammars to sets of strings, respectively. We can see each of these functions as augmenting a particular set of grammars with some structure. What a mathematical linguist will generally do in order to show this weak equivalence is to specify a way of converting a context-free grammar into a categorical grammar (and then the other way around), such that, when the language of either grammar is derived, it remains the same (Bar-Hillel, Gaifman & Shamir 1963; Pentus 1993). To say that this “way of converting” satisfies weak equivalence is to say it is a homomorphism in our

sense, and in fact a particular one: when we convert a CFG  $g$  into a CG  $F(g)$ , we require that the structure-defining function  $L_{\text{CFG}}$  be preserved. When we are given the task of writing the proof, we are given the constraints that  $F(L_{\text{CFG}}) = L_{\text{CG}}$ , which is to say that the “natural” definition of “string language” must be preserved, and that  $F(l) = l$  for any language in  $\mathcal{L}$ . So the demand of weak equivalence is one example of a demand for shared structure, and the challenge is to show that some structure (string language) can be preserved, in this case in a fairly exacting way (the strings need to be exactly the same).

Other kinds of equivalence are somewhat less stringent. One “strong equivalence” of CFGs and CGs says that, for a particular subset of the CFGs (Chomsky normal form CFGs), a conversion can be found to CGs that will preserve the derivation tree for any string. This is a stronger result in the sense that it preserves the action of a mapping with richer outputs (sets of string–tree pairs), but it is also weaker, in the sense that we must find a satisfactory way of converting our CFG derivations to CG derivations. The conversion is quite simple in this case, and follows from the conversion of the grammar (relabelling of nonterminals), but it is no longer an identity mapping. Nevertheless, the idea behind the challenge is the same: find a way of converting one grammar to another, such that some particular type of structure associated with the grammar (in this case, the set of sentence derivations) yields, if not precisely the same values, values which are “equivalent” in some meaningful way that we specify in advance (here, the tree has the same shape but only the corresponding, not the identical, labels on the nodes). We will look at this example in a bit more detail shortly.

To bring this back to our first general point: if one of the sets is a set of grammars allowed by some linguistic theory, and the other is the set of possible different linguistic

systems as implemented in the brain, then this is a very general way of seeing requirements that we as linguists place on our theories. The grammars we specify must at least have some structure in common with those that are in the brain. It can also be seen as a way of understanding the relation between different parts of the language faculty: a production system and a comprehension system may not work in the same way at all, and they might not be specifiable using the same information (they might have different “grammars,” broadly construed). For example, a speech perception system might require the acquisition of some auditory parameters, while a speech production system might require acquiring motor parameters. They might also operate over very different representations. But in some sense they follow the “same” grammar—in what sense? In the sense that they have some structure in common, be it an isomorphic yield of lexicon–surface pairs, a strong similarity between perceptual categories and the realization of their production counterparts, or merely some highly abstract similarity between the two perception and production inventories. That shared structure is the “competence” that these “performance” systems both instantiate. The first point, then: empirical demands on linguistic theories are assertions of structure sharing under homomorphism.

## 2.6.2 Notation and the structure of a grammar

Now, there is one very particular type of structure sharing that is relevant in the rest of this chapter. It is based on an idea that has been carried forward in some rough form from the evaluation measures of early generative grammar to the present day, but which has not had a precise sense since those early days: the form of the grammar itself matters,

not just the representations it works with, or the input–output mappings it countenances. In the current view, it is not the notation of the grammar per se that matters, but it is certainly something about the grammar itself, and not directly about the mappings it implies.

To continue our example in this context, consider a conversion between a Chomsky normal form CFG and a strongly equivalent CG. Here is a CNF grammar for  $a^n b^n$ :

$$(59) \quad \begin{array}{llll} S & \rightarrow & AB & B \rightarrow DZ & Z \rightarrow b \\ & & & D \rightarrow AB & B \rightarrow b \\ & & & & A \rightarrow a \end{array}$$

Now here is a CG for  $a^n b^n$ .

$$(60) \quad \begin{array}{ll} a: & S/B \quad b: \quad B \\ & D/B \quad B \backslash D \end{array}$$

The derivation under each grammar for  $aaabbb$  is shown in Figure 2.2. They are clearly isomorphic. In general, these two grammars will yield string–derivation pairs for which the derivations are isomorphic for a given string.<sup>12</sup>

---

<sup>12</sup>Precisely what the details of this “obvious” isomorphism are (what exactly  $F$  needs to say when applied to derivations for this correspondence to be interesting, and indeed how derivations are coded) is irrelevant here.

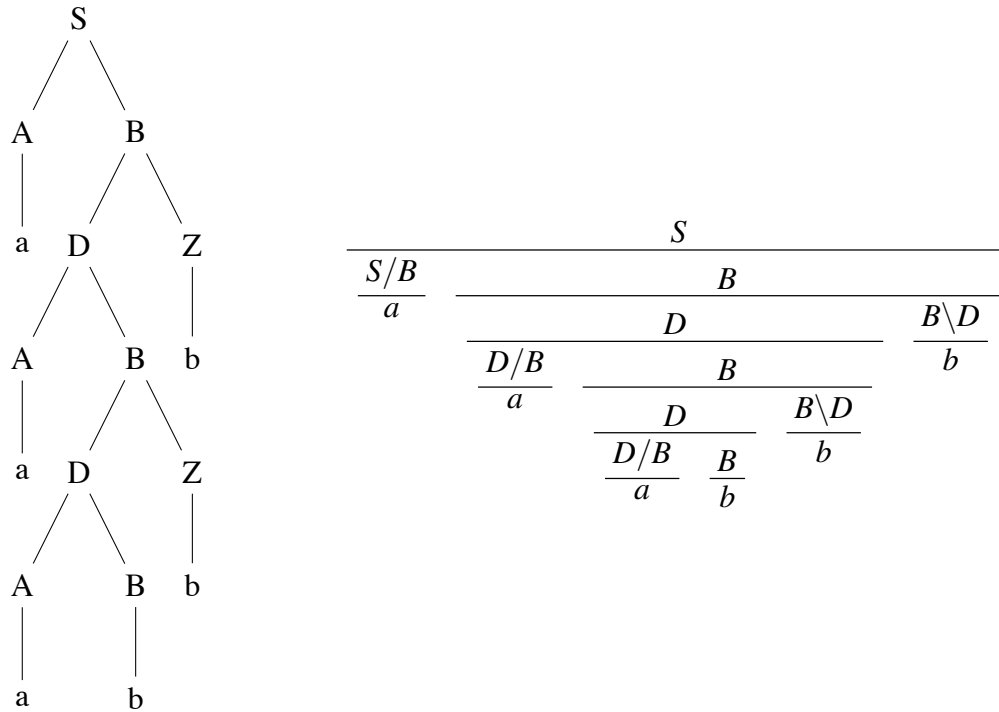


Figure 2.2: Derivations for the string  $aaabbb$  following to the CFG (left) and CG (right) given above. The CG proof is flipped vertically from its usual order (conclusions are above their premises) to emphasize the isomorphism.

We have already said that such a construction is possible, and in fact always possible in the case of CNF grammars. What we have not said is that there is another non-trivial property that these two grammars share. Cross off all but the first three lines from the CFG and cross off the second and fourth lines from the CG. We get these two new grammars:

$$(61) \quad \begin{aligned} S &\rightarrow AB \\ A &\rightarrow a \\ B &\rightarrow b \end{aligned}$$

$$(62) \quad \begin{aligned} a &: S/B \\ b &: B \end{aligned}$$

Both of these grammars specify the singleton language  $\{ab\}$ . Thus both of the original grammars “contain” a grammar for  $\{ab\}$ , notationally. Furthermore, for both

of the original grammars, the reader can verify that there is no way to remove any lines without yielding a grammar for  $\{ab\}$  or a grammar for  $\{\}$  (though in most cases with some redundant information in the form of useless productions or lexical entries; the latter are the grammars that cannot complete a yield/parse for any terminal strings, and/or contain no start symbol, and the minimal such grammar is the empty grammar). Thus both of the original grammars only contain grammars for  $a^n b^n$ ,  $\{ab\}$ , and  $\{\}$ . Since the grammars for  $\{ab\}$  also all contain grammars for  $\{\}$ , we can firmly state some relations in this small corner of the landscape of CFGs and CGs. For one thing,  $G_{\{\}} \prec G_{\{ab\}} \prec G_{a^n b^n}$ , where we mean the empty CFG, the CFG in (61), and the CFG in (59), respectively, and by  $\prec$  we mean, for now, “is a subpart of.” And it would seem, at least given these cases, that the CNF–CG conversion procedure respects this hierarchy:  $F(G_{\{\}}) \prec F(G_{\{ab\}}) \prec F(G_{a^n b^n})$ , where  $\prec$  is either the same or almost exactly the same relation across CFGs and CGs, depending on exactly how we formalize it. What’s more, each of these three grammars sets up an equivalence class, as each is the grammar yielded under “reduction” by the removal of useless productions or lexical entries from many others—in fact infinitely many others, and so, if we make it so that  $\prec$  evaluates containment not of  $G_1$  and  $G_2$  directly, but for the reduction of  $G_1$  and  $G_2$ , then we have now carved a path of shared structure through a highly restricted but infinitely large corner of the grammar landscape.

The important point here is that we have translated “different sorts of higher-order properties of the string that codes the grammar [may be] potentially meaningful [cognitively]” into the idea that there are separable pieces of meaningful information that make up a grammar, and these stand in a homomorphic relationship with separable pieces of meaningful information learned by the brain as the specification of (a particular part of)

the linguistic computation. In this case, we have pointed to a “subpart” relation as an example, which is intuitively easy to relate to the notion of “length.”

Before we break down the “subpart” relation into the more primitive structures that are directly implied by the statement of the grammar, let us dwell on it briefly to make sure that the notion of “structure” is clear, staying now on the CFG side. The grammar for  $\{ab\}$  given in (61) constitutes an addition of information to the empty grammar, and can also be refined in various ways, including the addition of another rule  $B \rightarrow c$  (this yields the language  $\{ab, ac\}$ ). This is shown in Figure 2.3.

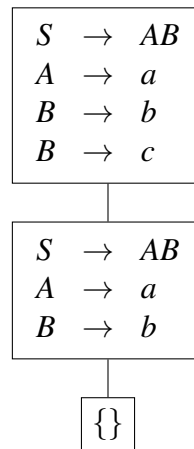


Figure 2.3: Diagram of subpart relations between three CFGs.

There is another grammar apart from our  $\{ab\}$  grammar which stands in the same relation with these two other grammars, namely the grammar for  $\{ac\}$  which has been added to the diagram in Figure 2.4.



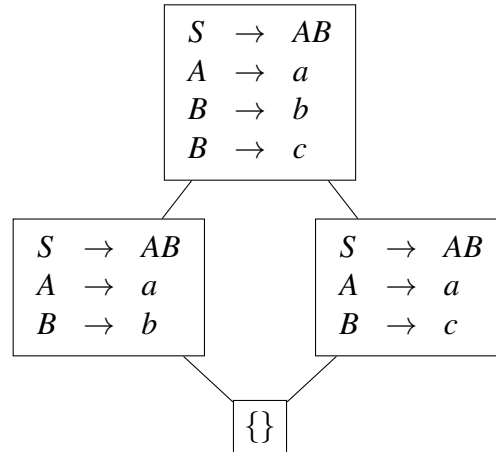


Figure 2.4: Diagram of subpart relations between four CFGs.

What these examples illustrate is that a context-free grammar takes the form of a collection of pieces of information of a particular kind (context-free rules), and this kind of specification gives rise to some structure: one can add or remove these pieces and get back different grammars, which induces a relational structure. If we had specified the same language, with the same derivational trees, in a different way, using “rules” that expand a tree not only by one level, but instead by substituting a more substantial part of a subtree, then we could have also had a grammar containing a single rule that would generated *aabb* with the same structure as assigned by this grammar, and the relational structure would have been different. The particular choice of grammar specification asserts a particular way of organizing the information used to specify the mapping, which is at least in part captured in the containment relations we have just illustrated. In fact, there need not be any subpart relations between grammars at all, depending on how we code the information in the grammar. Suppose we annotate every rule in the CFG, and every lexical entry in the CG, with the label entry  $[k]$  of a total of  $[n]$ . Every rule has such a label attached, and a collection of symbols is not allowed to be a grammar unless it has these labels correctly

marked; now, any time we remove lines, we get an illegal grammar—so now there can in fact be no subpart relations of the kind we have talked about. Again, the particular choice of grammar specification asserts a particular way of organizing the information used to specify the mapping. The next point of this section, again: we demand that the organizational scheme implied by a grammar formalism be real, in the sense that it is homomorphic with something in the brain.

Let us now spell this out this idea of an organizational scheme, or a structure, in a simple setting that is transparently extensible to any of the three particular examples of grammars we have discussed in this section. Suppose a grammar were a sequence of sequences of lexical items (imagine the right-hand side of a rewrite rule). Here is one way of organizing this information; we will translate it with some formal detail that introduces as few additional substantive assumptions as possible.

(63) “A Grammar contains a List of Lists of items from the Universal Lexicon”

$$\exists R. \exists L_R. \text{List}(L_R, R) \wedge \forall r. r \in R \rightarrow \exists S. \text{List}(r, S) \wedge \forall s. s \in S \rightarrow s \in UL$$

To properly fill out the examples that are to come, we also need the following condition:

$$\begin{aligned}
(64) \quad \text{List}(L, S) \rightarrow & \quad \exists s. s \in S \wedge \forall t. \neg \text{Prec}(L, t, s) \\
& \quad [\text{there is a head}] \\
& \wedge \quad \exists s. s \in S \wedge \forall t. \neg \text{Prec}(L, s, t) \\
& \quad [\text{there is a tail}] \\
& \wedge \quad \exists s. \forall u. [\forall t. \neg \text{Prec}(L, t, u)] \rightarrow u = s \\
& \quad [\text{head unique}] \\
& \wedge \quad \exists s. \forall u. [\forall t. \neg \text{Prec}(L, u, t)] \rightarrow u = s \\
& \quad [\text{tail unique}] \\
& \wedge \quad \forall s. \exists t. [t \in S \wedge \text{Prec}(L, t, s)] \rightarrow [\forall u. \text{Prec}(L, u, s) \rightarrow u = t] \\
& \quad [\text{precedents unique}] \\
& \wedge \quad \forall s. \exists t. [t \in S \wedge \text{Prec}(L, s, t)] \rightarrow [\forall u. \text{Prec}(L, s, u) \rightarrow u = t] \\
& \quad [\text{successors unique}] \\
& \wedge \quad \forall s. \forall t. \text{Prec}(L, s, t) \rightarrow \neg \text{Prec}(L, t, s) \\
& \quad [\text{precedence asymmetric}]
\end{aligned}$$

Simply asserting the sentence (63) by itself is not enough to specify a grammar of this kind. The identity of the lexical items and what exactly the orders (which are associated with particular Lists) are are properties of the objects that satisfy the formula, not of the formula itself. Given a finite collection of objects, the sentence (63) asserts the constraints on a logical structure that makes it into a grammar. In particular, suppose we have some sentence  $P = \exists a_1 \dots \exists a_k. \Phi(a_1, \dots, a_k)$ , where  $\Phi$  is quantifier-free, which has (63) as a logical consequence (under any interpretation that satisfies our sentence, (63) will also be satisfied). Call  $P$  a grammatical description. If our collection of objects  $D_P$  can be used as the domain for some model  $\mathcal{M}_P \models P$  (with basic relations like elementhood fixed in the model in a standard way), then  $\langle P, D_P \rangle$  is a grammar. We may call  $D_P$  the gram-

mathematical content for the sake of having a label, although there is surely content in  $P$  also (after all, there is more than one possible  $P$ ). In this sense, the sentence (63) induces an “organizational structure” associated with the set of grammars.

We can then assert interesting derived structures. For example, suppose we have some grammatical descriptions  $S$  and  $T$ . If any grammatical content that satisfies  $S$  also satisfies  $T$ , then  $T$  is a grammatical consequence of  $S$ . Now suppose we take a model that satisfies  $S$ , called  $\mathcal{M}_S$ , and remove some objects from the domain or some relations from the model, as well as references to the removed objects among the functions and relations of the model, to obtain a new model  $\mathcal{M}_T$ . If  $\mathcal{M}_T \models T$  but  $\mathcal{M}_T \not\models S$ , and if the set of atomic domain elements of the model  $\mathcal{M}_S$  (here we will take any elements of  $UL$  to be atomic) is minimal while both satisfying  $S$  and containing the elements of  $\mathcal{M}_T$ , then we might say that  $\langle T, \mathcal{M}_T \rangle$  is a subpart of  $\langle S, \mathcal{M}_S \rangle$ , in a particular sense. (The atomic minimality condition ensures that the difference between the nested models is not such that the elements of the smaller one are no longer useful in the larger model, in which case one could simply add many elements to model the entire new sentence; the condition rests on the notion that in that case we could construct a more complex sentence still which would make use of the newly useless elements.)

Here is an example. Consider the grammar  $[[A, B], [a], [b], [c]]$  (remember that in this setting, the “rules” are ordered). Seen as a set of objects, this must be at least (in order to model a grammar description):

(65) Objects  $A, B, a, b, c$ , which have the property that they are elements of the object

$UL$  (a constant whose makeup is fixed); our use of labels here is arbitrary but is

meant to make clear that the objects are distinct

- (66) Objects we might label  $\{A, B\}, \{a\}, \{b\}, \{c\}$  to make clear how the “element” relation holds with respect to the objects just mentioned
- (67) Objects we might label  $[A, B], [a], [b], [c]$  to make clear that the List property holds of them, and which of the objects just mentioned they are lists over, and to indicate what Prec relations hold
- (68) An object  $\{[A, B], [a], [b], [c]\}$  (given this label by us as only a mnemonic for the relations it participates in, as before)
- (69) An object  $[[A, B], [a], [b], [c]]$  (as before)

A sentence that describes a grammar that could be modelled under this domain is as fol-

lows:

$$\begin{aligned}
(70) \quad & \exists R, L_R, r_1, \dots, r_4, s_1, \dots, s_4, l_1, \dots, l_5. \\
& \text{List}(L_R, R) \wedge r_1 \in R \wedge \dots \wedge r_4 \in R \\
& \wedge \text{List}(r_1, s_1) \wedge \dots \wedge \text{List}(r_4, s_4) \\
& \wedge l_1 \in s_1 \wedge l_1 \in UL \wedge l_2 \in s_1 \wedge l_2 \in UL \\
& \wedge l_3 \in s_2 \wedge l_3 \in UL \\
& \wedge l_4 \in s_3 \wedge l_4 \in UL \\
& \wedge l_5 \in s_4 \wedge l_5 \in UL \\
& \wedge \text{Prec}(L_R, r_1, r_2) \wedge \text{Prec}(L_R, r_2, r_3) \wedge \text{Prec}(L_R, r_3, r_4) \\
& \wedge \neg \text{Prec}(L_R, r_1, r_1) \wedge \neg \text{Prec}(L_R, r_2, r_1) \\
& \wedge \neg \text{Prec}(L_R, r_3, r_1) \wedge \neg \text{Prec}(L_R, r_4, r_1) \\
& \wedge \neg \text{Prec}(L_R, r_4, r_4) \wedge \neg \text{Prec}(L_R, r_4, r_1) \\
& \wedge \neg \text{Prec}(L_R, r_4, r_2) \wedge \neg \text{Prec}(L_R, r_4, r_3) \\
& \wedge \neg \text{Prec}(L_R, r_2, r_2) \wedge \neg \text{Prec}(L_R, r_3, r_2) \wedge \neg \text{Prec}(L_R, r_4, r_2) \\
& \wedge \neg \text{Prec}(L_R, r_1, r_3) \wedge \neg \text{Prec}(L_R, r_3, r_3) \wedge \neg \text{Prec}(L_R, r_4, r_3) \\
& \wedge \neg \text{Prec}(L_R, r_1, r_4) \wedge \neg \text{Prec}(L_R, r_2, r_4) \wedge \neg \text{Prec}(L_R, r_4, r_4) \\
& \wedge \text{Prec}(r_1, l_1, l_2) \\
& \wedge \neg \text{Prec}(r_1, l_1, l_1) \wedge \neg \text{Prec}(r_1, l_2, l_1) \wedge \neg \text{Prec}(r_1, l_2, l_2)
\end{aligned}$$

Now, another grammatical description could be satisfied by the same model if it simply failed to introduce require any new distinct objects or inconsistent orderings (both distinctness and ordering here are demanded only by Prec relations and the conditions in (63) that

govern them). For example, one could remove  $r_4$  and all references to it from the sentence; or one could remove  $l_5$  and all references to it; or both. In any case, the same model would suffice unchanged to satisfy the new sentence—the absence of any assertions demanding the existence of an element of the domain does not mean that that element does not exist. This would be true for any other model of the larger sentence, not just this one. The new sentence would be a grammatical consequence of the old one.

Suppose we remove  $r_4$ . In a sense, we have a grammatical description of something we might call  $[[A, B], [a], [b]]$  (regardless of whether we include the object corresponding to  $[c]$  in the model). However, this is somewhat misleading, as the description would be satisfied with alternate objects in the domain: we could have just as easily written  $[[C, D], [e], [f]]$ , because there is also a model for the same grammatical description that has that object at the top level. If we put the demand on the model that its domain be a proper subset of the one outlined above, however, with none of the relations changed except to remove references to the given elements, then the only way to continue to satisfy all the requirements of the new sentence in a minimal way is to remove the object we called  $[c]$  (we have already changed the top-level list to something we would more naturally label  $[[A, B], [a], [b]]$  by updating the relations in the model to remove references to  $[c]$ ).<sup>13</sup>

This comports with our intuition that the resulting grammars are nested, with one being a

---

<sup>13</sup>We could remove  $c$  and  $\{c\}$ , in fact, and make the model still “more minimal”—not because there are no longer any conjuncts in the sentence that used to be satisfied by virtue of their inclusion ( $l_5 \in s_4$  is still in the sentence) but because in fact they never needed to be included in the first place. While  $r_4$  needs to be distinct from  $r_1, \dots, r_3$  because the four have conflicting precedence requirements,  $l_5$  and  $s_4$  have no precedence requirements and the relevant clauses could be satisfied by other domain objects, such as  $b$  and  $\{b\}$ . In fact, only two lexical items are needed in the model which is truly minimal. This seems to demand that either that the universal lexicon be treated as a set of properties, or else that the notion “subpart” is actually only relevant for this subset of the grammars that are truly minimal. This seems to amount to a technical detail and does not change the basic point that a grammatical framework is a set of constraints on the organization of grammatical intuition, and that these constraints can give rise to subpart relations, so I will not pursue it further here.

subpart of the other.<sup>14</sup>

In short: we have given a rough outline spelling out the types of constraining information that are implicit in a grammatical formalism, and shown that these constraints are a structure on collections of information over which further structure can be defined.

It should be pointed out that some theories of grammar could be formulated so that every grammar has the same grammatical description, and differs only in content. For example, in standard monostratal Optimality Theory, a grammar consists of an order on a fixed, universal set of constraints. No organization is demanded except to say that the only available relation is something like *Prec*, or that there is a mapping from constraints to the integers or some other set with a fixed order. Presumably, any grammatical description for such a grammar is a grammatical consequence of any other; similarly for the Principles and Parameters approach. We have already discussed an issue that—as we are about to show—can be reduced to the same thing, namely, the irrelevance of a notion of simplicity for these types of grammars. The conclusion we reached above was that the apparent fixed length and uniform information content of these grammars is more limited than we might think naively, because they need to be supplemented by other kinds of variable-length information that interact with this fixed-length information.

In particular, we briefly touched on the problems of picking out a subset of the set of logically possible constraints, and of picking out a subset of the set of logically possible lexical items, both generally understood to be real learning problems (that is, things that

---

<sup>14</sup>It should be noted that there are other, different cases of grammatical consequence. For example,  $T$  is a grammatical consequence of  $S$  for  $T = \exists x.P(x) \vee Q(x)$ ,  $S = \exists x.Q(x)$ , or  $T = \exists x.P(x)$ ,  $S = \exists x.P(x) \wedge Q(x)$ . In the first case, as in the subpart case, the consequent sentence tolerates a whole class of models that  $S$  does not, but, unlike in the subpart case, this is not due to the addition of an object or relation to the model in  $S$ . In the second case, an additional relation, but not an additional object, is demanded in the model by  $S$ .



need to be specified somewhere). To take a concrete example, the difference between the theory that the set of universal constraints is finite and the theory that it is infinite is that in addition to the order (or List) over the constraint set, the grammar also needs to specify which subset of the constraint set the order is over, and so needs to predicate something of each element of that subset.<sup>15</sup> A grammatical description for one such possible grammar would assert the existence of more elements than another, and thus it would no longer be the case that all grammatical descriptions were consequences of all others. That would allow for the kind of higher-order structure we are discussing here to be asserted, and, as we will see, that, in turn, provides a natural way for the Bayesian Occam's Razor to apply automatically to some inferences in these frameworks.

### 2.6.3 Relating grammars to priors in an optimal way

Here is an idea about how we could construct priors from grammars. It is very weak, but it is enough to get the Bayesian Occam's Razor to hold:

(71) Optimal Measure Principle. If  $T$  is a grammatical consequence of  $S$  then the prior

evaluation measure  $\Pr[\cdot|S]$  is similar to  $\Pr[\cdot|T]$  for  $M_S^* \subseteq \{\mathcal{M}|\mathcal{M} \models S\}$ ,

$$M_T^* \subseteq \{\mathcal{M}|\mathcal{M} \models T\} \setminus \{\mathcal{M}|\mathcal{M} \models S\}.$$

---

<sup>15</sup>It is worth emphasizing that, while we have made use of domain elements that have the properties of complex objects such as sets and lists, they were actually contentless placeholders, as they only had these properties in virtue of relations external to themselves; only the atoms (in the above, those domain elements which are part of the universal lexicon) can be thought to have intrinsic content. If we had instead had a set, such as the set of language-specific constraints or lexical items, as a contentful element of the domain, rather than having the elements in the domain and relating them to a placeholder object, then we would be asserting that the grammar is one-dimensional, as opposed to  $|\mathcal{U}|$ -dimensional (in this case, infinite-dimensional). That is fine as an alternate structure for the grammar, but it needs to be kept in mind that the question is an empirical one if we demand that the grammatical description be shared structure with the brain, which is to say if we take the strong but axiomatic demand of this whole chapter seriously.

If the grammatical description stands in for the framework parameter  $\omega$ , this asserts that FCP holds for grammars. In particular,  $T$  represents the structure of a grammar which is just “like” another, described by  $S$ , except that it is satisfied with some models ( $M_T \setminus M_S$ ) that  $S$  is not—in the cases we examined above,  $T$  required “less” information in the sense of a smaller domain. The meaning is that there must be a bijection between some of the “less” and “more informative” models so that  $\Pr[\cdot | M_S^*, S] = \Pr[\cdot | M_T^*, T]$ . These conditions on the prior imply that, at least for the shared subsets, the prior distribution on information within those subsets is independent of the choice of structure. We will discuss concrete examples in more detail shortly, but for the moment think of the presence or absence of an additional rule: for at least some such additional rule, the learner’s biases on the contents of the rest of the grammar do not change simply because that rule is present or absent.<sup>16</sup>

This condition is extremely weak, however, and, although grammatical consequence clearly gives some structure across which we can now posit mappings, and furthermore clearly gives rise to at least some mappings between  $M_T^*$  and  $M_S^*$  for which the model from the first class represents a “less complex” grammar than the grammar from the second, it is far from clear that all such cases will be like this. We can place a stronger condition by adding additional structure preservation to the condition for similarity:

(72) Asymmetric similarity over subpart models: Two prior distributions  $\Pr_1$  and  $\Pr_2$

over two sets of models  $\Theta_1, \Theta_2$  for sentences  $T, S$  respectively are similar for some

$A_1 \subseteq \Theta_1, A_2 \subseteq \Theta_2$  if there is a bijection  $f : A_1 \rightarrow A_2$  such that, for all  $s_M \subseteq A_1$ ,

---

<sup>16</sup>Actually, it is even slightly weaker than this: we could permute the possible “smaller” grammars until their prior aligns with the prior on that part of the grammar in the “larger” structure. There is no guarantee that the priors can ever be made to align; but what it does show is that the BOR will hold under a wide range of conditions.

$\Pr_1 [s_M | A_1] = \Pr_2 [f(s_M) | A_2]$ , each domain element and relation of  $s_M$  is also present in  $f(s_M)$ , and  $f(s_M)$  is atom-minimal for a domain element which both satisfies  $S$  and contain  $s_M$ .

Now it is clear that, if a model for the consequent sentence does not satisfy the antecedent, it is due to some missing element, because the presence of this element on the other side of the bijection is enough to guarantee satisfaction of the antecedent. In the case where  $A_1$  is the entirety of  $M_T \setminus M_S$ , this stronger notion of similarity is enough to guarantee that the failure of certain models in  $M_T \setminus M_S$  to model  $S$  is entirely due to this difference in complexity; in this case, the BOR must hold in favor of the sentence  $T$ , all things being equal.

Although this condition is still fairly weak, to the extent that it holds it falls intuitively into the class of “optimal” principles sought out under the Minimalist program of Chomsky 1995: although the sense of “optimal” is somewhat vague, the intuition about the OMP, where it holds, is that the information in the prior “follows from the structure” of a particular grammar to a large extent, because non-changes to some sub-part of the structure of the grammar track non-changes in the prior over the contents of that substructure.

#### 2.6.4 Example: deriving a symbol-counting evaluation measure

Recall the prior evaluation measure of Chomsky & Halle 1968 discussed above: The ‘value’ of a sequence of rules is the reciprocal of the number of symbols in its minimal representation. Putting aside the notion of “minimal representation” (which allows for the

collapsing of environments using brackets and so forth), we can simply take the measure to evaluate a particular representation of a particular grammar. First note that this will never sum to one, nor any other finite quantity (the harmonic series is not convergent), and thus cannot be a probability measure. However, it can be weakened to merely some function decreasing in the number of symbols in the grammar, with no real consequence for any of its uses in the literature.

Although it is now clear that the OMP will guarantee such preferences in the prior in the fairly general case where the two grammars stand in a consequence relation for which the antecedent grammar description can be satisfied by adding to some model for the consequent one—in which case the antecedent grammar is clearly “more complex”—the full import of the similarity condition and the effect of the likelihood has not been explored intuitively in the grammatical context.

Begin with the similarity condition: this says that the addition of some clause necessitating an additional domain element or relation in the model (which would presumably be represented as additional symbols in a meaningful grammar notation) does not change the relative preferences for existing elements of the grammar, and under the strong version just outlined, this must be the case across the entirety of the set of possible models. Suppose we add a rule to the grammar  $A \rightarrow B/C - D$ , which we would perhaps spell out formally as a List, following the above,  $[[[A], [B], [C], [D]]]$ , now  $[[[A], [B], [C], [D]], [[E], [F], [G], [H]]]$ . The other possible models satisfying the description of the first grammar contain alternate lexical items, and may also contain extraneous elements which would nevertheless not serve to satisfy the description of the second grammar. Focusing on the choice of lexical items, the distribution over the choice of items in  $A \rightarrow B/C - D$  must be the same regard-

less of the presence or absence of the other rule. (Recall that this is the prior distribution, so no consideration of consistency with the data is needed.)

The only thing worth pointing out about the condition on the likelihood is that it may be gradient, unlike the classical evaluation measure, handling, for example, the aggregate consistency with all the available data. The precise specification of the prior will yield a precise trading relation between the goodness of fit made available by the various different models across the two grammatical descriptions, on the one hand, and the BOR scaling factor, on the other.

## 2.7 Discussion

The goal of this analysis has been to highlight the circumstances under which Bayesian inference for grammars will give rise to a simplicity bias. We return now to questions about the paper by PTR, in light of these new tools. Why does the PTR model work? Does it have to do with the Bayesian Occam's Razor? Does the PTR prior obey the OMP?

We now know a good way to identify cases of BOR: look for hyperparameters which imply a change in the “size” (in terms of the uncertainty in the prior measure) of the set of possibilities for other parameters in the specification of a grammar; the clear cases will be those for which these other parameters' distributions are independent of the choice of hyperparameter. The PTR prior is riddled with these: there is a (trivial) distribution on the selection “regular-only versus CFG”; there is a distribution on the total number of nonterminals; there is a distribution on the number of productions; and there is a distribution on the number of items on the right-hand side, for a given rule. In each case, the choice does

not affect the prior on the dependent parameters at all. Note that it is independently true that these hyperparameter distributions themselves contain biases for smaller grammars (chiefly because they are distributions over the positive integers and decay as they go to infinity); but this is actually unrelated to the BOR. There would be a preference for smaller grammars regardless, simply because of the structure of the prior. Notice, however, that in the case of “regular-only versus CFG,” the “smaller” set in our “less diffuse” sense is clearly the set of right-regular grammars. The model prefers proper CFGs in spite of, not because of, the BOR in this case.

As discussed earlier, and as pointed out by PTR, the problem with right-regular grammars is that they must yield bad analyses for many sentences, even where they fit well. The analyses might be bad because they need to make use of productions which give away too much probability to unattested sentences (where the probabilities in the grammars have a high degree of uncertainty about which productions should apply), thus dispreferred by restrictiveness; or they might be bad because they are too large, which would lead to problems with simplicity and restrictiveness both. PTR’s results are that the likelihood prefers regular grammars, while the posterior prefers CFGs, thus pointing to some combination of the BOR and the other biases in their prior. The increase in size (which would be dispreferred by both types of biases) is in the number of rules, which must increase to obtain a similar fit in the regular grammars, as PTR point out.

PTR’s result thus really does follow from a plausibly “domain-general” effect, that of the very general law of inference we call the BOR, at least to some degree. However, as pointed out above, the suggestion that particular choice of hypothesis space is one that stands in well for some plausible set of hypotheses available to “general cognition” any

better than it stands in for one which is specific to language has no real basis.

What is to be done with the Bayesian Occam's Razor? Of course, it will be embedded in most hierarchical Bayesian models we use to account for language acquisition; however, there is more that can be said about this. In particular, with the knowledge that “flat” grammars (classical OT and P&P) need not be equipped with a notion of simplicity, while structured grammars must, because of BOR, it would be reasonable to expect that flat and structured grammars ought to make systematically different predictions about acquisition and historical change, which we should be able to extract reasonably clearly just by sketching out the predictions in particular cases informally. A thorough review of attributions of various historical changes to simplicity throughout the literature would be a reasonable place to start.

With such evidence in hand, one can then use Bayesian statistics not only as a theory of language acquisition per se, but also as a meta-theory: a theory about what linguistic theories do and do not predict, which can be used to derive predictions about those theories (in this case, about the behavior of a learner), and then compare them.

Finally, it is worth discussing the ontological status of the sorts of formal constraints we used to translate the notation of grammars into the organization of certain pieces of information. In particular, while it seems quite clear that some of what learners do inference over is indeed “organizational,” which is to say, referring to the structure of the grammar (its size and shape), the utility of such information outside acquisition is somewhat difficult to understand. If, say, parsing makes reference only one grammar at a time—which is to say, one specification of the necessary information necessary to get a human parser to work—then why should that information be structured in any way that resembles the

organization of that information for the learner? This is somewhat relevant, because the more “natural” the organizational principles of grammar, the more “optimal” the OMP seems to be.

There is one reason that I can think of, which is that, if the length of grammars is unbounded, then in fact there will be some possible grammars for which only certain subsets of the information they contain will be able to be accessed at a given time by any device whatsoever with a finite memory. This is definitely true for the lexicon anyway—it would surely be impossible to make simultaneous use of information about large numbers of lexical items—and it would therefore be quite interesting if some organizational structure of the lexicon (say, as uncovered by lexical access times) were a place the BOR could hang; we could then investigate learners’ preferences to see whether they did indeed track this structure.

## 2.8 Conclusion

In this chapter, I have presented a detailed explanation of what it means to talk about the Bayesian Occam’s Razor which is sometimes referred to in the Bayesian literature. I have outlined the action of this law of inference, and certain general conditions under which it will hold. I have then discussed what it means for a grammar to itself have structure (rather than merely assign structure), and claimed that this structure is what is really being referred to when the “notation” of a grammar is taken to embed some empirical claims. Using a relatively neutral example of how such structure could be spelled out, I have developed what it would mean to apply the Bayesian Occam’s Razor to a structured



grammar. Finally, I have argued that the BOR makes Bayesian statistics a particularly useful tool for investigating various theoretical and theory-comparison questions from new angles.

## Chapter 3: Modelling allophone learning

The first question I ask myself when something doesn't seem to be beautiful is why do I think it's not beautiful. And very shortly you discover that there is no reason.

— Attributed to John Cage

### 3.1 Categories and transformations

This chapter proposes Bayesian models for learning segmental categories and allophonic processes, two crucial parts of linguistic cognition studied in phonology. The models are based on a new idea about what it means to be a context-dependent “phonetic rule,” namely that there is an “addition” operation in a gradient phonetic space, and each separate effect of context is its own addition. The conjecture is furthermore put forth that all cases of allophony are phonetic rules in this sense. I use the learning models to argue that such a model is feasible, and that something like this might even be crucial to learning phonetic categories.

#### 3.1.1 Empirical review

Phonology investigates the cognitive systems involved in producing and recognizing speech: the auditory system as it applies to speech, the motor system, the system of lex-

ical memory that underlies the ability to store and recall the forms of words. There is also a connection to traditional grammar, which, among other things, tries to describe patterns in how different sounds are pronounced in different contexts in a particular language. Once we know what these patterns are, certain crucial facts about them constrain our understanding of the cognitive mapping that converts lexical representations to pronunciations. In Chapter 1, I reviewed the standard assumptions about what is in the lexical memory system: lexical memory for a form consists of a finite sequence of segments reflecting the sequence of sounds, and each segment can be classified as a member of some discrete set (the inventory). Well-understood patterns in pronunciation seem to respect segments and inventories, and we use the fact that they do to support theories about what information is in the lexicon. The linguistic patterns in question are the processes discussed in Chapter 1 (both neutralizing and strictly allophonic). The idea behind saying that processes respect segmentation is that processes do not have effects that arbitrarily subdivide words; the idea behind saying that processes respect the discrete-valued nature of segments is that, when a segment changes from one to another, the resulting segment is pronounced just like other instances of that segment which are coded lexically, or which result from other processes (that is, the resulting phonetic realizations are statistically the same as for some other category, all other things being equal). If the patterns are understood as changes in pronunciation that are carried out in a mapping from lexical to pronounced form called the phonological grammar, then discrete-valued segments can be seen as crucial parts of the computation of this mapping, and segments must have some cognitive status in the lexical representations and the grammar that manipulates them. The goal of research in phonology has been to find a way of formulating these mappings that satisfies the usual

requirements for a linguistic theory: they must share some structure with the way the brain does it.

Beyond the fact that it gives rise to linguistic patterns in pronunciation which generalize in particular ways, there are other sources of evidence about this division of lexical representations into segments, and the classification of segments into discrete categories. Speech perception experiments often ask speakers to identify many slight variations on a sound, along some acoustic continuum, (“is it ee as in bee or ih as in bit?”), and then test the ability of listeners to discriminate between these small changes in pronunciation. Discrimination ability tracks the identification curves, meaning that, as judgments become clearer about which segmental category a sound belongs to, people’s ability to tell small differences apart gets worse, a phenomenon which is referred to as categorical perception or the perceptual magnet effect (Abramson & Lisker 1970, Pisoni 1973, Kuhl 1991). This suggests that discrete classes of segments have some special cognitive status apart from just being convenient ways of labelling the stimuli in the task; this idea that there are categories—equivalence classes of sounds—follows the usual understanding of how the phonological system works, in particular, the idea that phonemes form categories. The reasoning is not airtight, as the association of categorical perception with discrete cognitive categories is only one explanation. Another would be that perception is warped in a way that simply happens to align with the identification curves, perhaps because the prototypical pronunciations relatively dense statistical clusters in the input. Nevertheless, one does not need to do an experiment to see that processes that implicate phonological representations have categorical effects: when listeners hear a sequence that is either pit or bit, they surely either believe it is one word or the other—or assign some probability

to either—but not some interpolated word; so at some point in the process of recognizing speech, the encoding of a sound needs to have some information about whether it is codes for one of these words or another.

The idea that there are segments in the first place says that words are broken down into sequences—it is not just a matter of recognizing a sound as coding either bat or bad, rather, that decision is made up of smaller decisions about b–a–t or b–a–d. Again, linguistic patterns seem to operate over temporal sub-syllabic chunks, but a skeptic could argue that this does not require that these temporal chunks have any cognitive status. Independent evidence for such chunks comes from speech error research: many speech errors seem to be substitutions of entire segments, like [blejkfrud] for [brejkflud], “brake fluid,” and [frto] for [frto] “fish grotto” (Fromkin 1973, Dell 1986, Frisch & Wright 2002). Wan & Jaeger 1998 report that this is true even for Mandarin speakers in Taiwan, where, at least at the time their experiment was done, the phonics instruction in school was done entirely using a quasisyllabic orthography called Zhuyin (bopomofo), not the alphabetic pinyin system used on the mainland, meaning that there is no chance that the speakers were relying on some extra-linguistic representation of the written form, (as has sometimes been claimed to be responsible for segment effects in other languages), because speakers sometimes make segmental errors within what would be single Zhuyin symbols. Thus it appears that the division of lexical stored forms into segments, and the discrete classification of those chunks into phonemes, shows some cognitive effects apart from the fact that there are phonological alternations that obey this discretization.

As for the processes themselves, the experimental literature focuses on the strictly allophonic process discussed in Chapter 1. Strictly allophonic processes output segments

that do not appear anywhere apart from in the output of these processes, like the example discussed there of Spanish [bota]/[laβota]. The traditional understanding is that the distinction is not coded in the lexicon, because it does not need to be, and it seems to be the case that listeners are worse at perceiving the difference (for Spanish, see Boomer-shine, Hall, Hume & Johnson 2008). There are two types of results: the first shows that strictly allophonic pairs of sounds are not distinguishable or less distinguishable in perception; the other shows that, while two sounds might be distinguishable sometimes, in a form where an allophonic process applies, listeners are worse at telling the resulting allophonic pronunciation of the sound apart from the other one. Each may true for different allophonic pairs, depending on the status of those sounds in the language. The first approach tries to separate out the effects of allophony on the perception of segmental categories per se from the effects of perceiving in the presence of an allophonic alternation; Kazanina, Phillips & Idsardi 2006 showed that, even outside of the environment where the allophonic change occurs, Korean speakers, for whom [d] and [t] are strict allophones, cannot tell these sounds apart (between sonorants the voicing on stops changes, but the result, a sound like [d], is not a kind of voicing that can ever code for a distinct word from the one with a corresponding [t]; it is in other languages, like Russian, and Russian speakers, naturally, discriminate these sounds perfectly well). The second shows that, even if the segments themselves are in some cases distinguishable, allophonic processes still have a detectable impact on perception: Peperkamp, Pettinato & Dupoux 2002 showed that French speakers had no problem discriminating the voiced–voiceless pair [a]/[aχ] in isolation, but found it very difficult to discriminate [azo]/[aχzo], (they told them that they were hearing a new language and the two syllables were separate words), because there is

a process of voicing assimilation in this environment; there is no lexical contrast between [ ]/[χ] at all in French, so the process is truly allophonic, but voicing is contrastive for other fricatives in French, which presumably explains the results in isolation.

The facts about perception do support some cognitive status for allophony and the lack of contrast it induces, and the traditional approach, documented in Chapter 1, is to say that the stored forms do not code the allophonic distinction, (for example, they code allophonic [χ] as if it were just another [ ]), and then it gets added by the phonological grammar. However, there is another approach which says that allophonic processes are not true processes, in the sense of changes from lexical form to pronounced form, carried out by the grammar. This has become standard in much recent phonological theory: a principle called lexicon optimization is thought to determine what happens when lexical forms need to be stored in memory, and this principle essentially says that that learners will store underlying forms in a way that makes them as similar as possible to what was perceived (see Prince & Smolensky 2004 for an explanation of why the way Optimality Theory works makes it tempting to invoke such a principle). This means that, unless there are independent examples of a morpheme appearing in its pre-allophonic form—unless a morpheme is pronounced as [a] in one form, but the pronunciation of that same morpheme changes when some ending is added, to make something like [aχso]—then the lexical form will be just like the perceived pronunciation, and so the stored form is usually thought to code all the non-contrastive allophonic segments in many cases. The result is that allophonic “processes” are no longer grammatical changes in this theory, but merely static grammatical knowledge of cooccurrence restrictions (phonotactics: see below). By itself, this does not predict that listeners’ perception should be affected by allophonic patterns,

although it does not rule it out; however, the results from speech perception studies are somewhat troubling for this perspective. After all, if listeners do not reliably discriminate allophonically related segments, does this also mean that the distinction is not made anywhere up the processing stream—and, crucially, is it visible to the phonological grammar? If not, then the right way of thinking of the cognitive status of allophonic processes is not as co-occurrence restrictions, because in that case there would simply be no way to state such a grammatical restriction.

In any case, listeners unpack the allophonic processes and correctly recognize the words they affect, and learn, as speakers, to reproduce the allophonic pattern. Thus, something must be said not only about how phonetic categories are learned, but also about how the allophonic patterns themselves are learned (along with everything else). I will save the discussion of why allophonic processes deserve this special attention for Chapters 4 and 5. I turn now to some learning models for each of these things, before introducing a new hypothesis about how allophony works, and reviewing a new learning model for both categories and allophonic processes.

### 3.1.2 Computational and mathematical models

Imagine the following procedure:

1. Choose one of a finite number of items stochastically (maybe with some bias)
2. The items are each associated with a probability distribution over observables—say, sets of vowel formant measurements, or anything at all, like colors on a color wheel—so choose a particular observable in proportion to how likely the selected



probability distribution says it ought to be

We start with a probability distribution over categories, (items), a finite set of possible values for some variable we could call  $z$ ; we select from that distribution; we use the outcome to change how we (stochastically) select something else, call it  $y$ , an observable. This is a generative model for what is called a mixture model. It is called this because if we were to ask what the relative probabilities of different observables are, we would not be able to give just one answer: we would have to outline all the possibilities, one for each of the items we might have drawn—a mixture of different distributions.

Now imagine doing the process in reverse: given some observable, choose one of the items. The criterion is to choose the “best.” In our simple Bayesian formulation, we will say that what this means is to maximize the posterior probability of  $z$  given  $y$ ; we will use Bayes’ Rule to compute this (see Chapters 1 and 2); but this is not the only way of making this decision. There are many ways of doing inference back to the category selection in this way—but a phonetic category system always consists of a set of categories, each of which gives rise to a different perceptual map, and so the problem of perceiving speech sounds can necessarily be seen as having the same abstract structure as inverting a generative mixture model. Furthermore, the problem of learning the categories necessarily has the following structure: learn the phonetic maps associated with each possible phonetic map; learn the perceptual bias for different categories, if there is any. Then, so long as the preferences in the phonetic map and in the perceptual map follow probability theory, the phonetic category learning problem must be the problem of learning a mixture model.

Previous research on phonetic category learning often uses standard statistical methods for fitting mixture models. The perceptual maps are generally multivariate Gaussian distributions (out of convenience), which, in our figures, will appear as in Figure 3.1. Figure 3.1 shows a mixture of three two-dimensional Gaussian distributions. Multidimensional Gaussian distributions generalize the single-variable Gaussian distribution—a symmetrical probability distribution based on a “sum of squared error” computation, with a location parameter  $\mu$  setting the center, and a scale parameter  $\sigma^2$ , the variance, setting how quickly the probability falls off away from the center—to multiple dimensions. The center becomes a  $p$ -dimensional vector, and the scale becomes a  $p$ -by- $p$  matrix, listing not only the variance on the  $p$  dimensions, but also their covariance (unscaled correlation). The set of all observations with probability no less than some fixed  $\pi$  is an ellipsoid which is aligned with the axes if the variables on the  $p$  dimensions are not correlated, and otherwise has some rotation in proportion to the degree of correlation; it is the covariance matrix that sets this size and shape, while the location parameter is responsible for where the ellipse is centered.

Previous phonetic category learning models employing Gaussian mixtures include: de Boer & Kuhl 2003 (long English corner vowels, /i/, /e/, /u/); Vallabha, McClelland, Pons, Werker & Amano 2007 (four English and four Japanese vowels, /i/ and /e/ and their short—for English, lax—counterparts; their data was resampled from Gaussians; they also used a mixture of nonparametric category distributions in place of Gaussians in a second experiments); McMurray, Aslin & Toscano 2009 (voice onset times for English stops; data was resampled from Gaussians); Feldman, Griffiths & Morgan 2009 (all English vowels, taken from Hillenbrand, Getty, Clark & Wheeler 1995; data was resampled

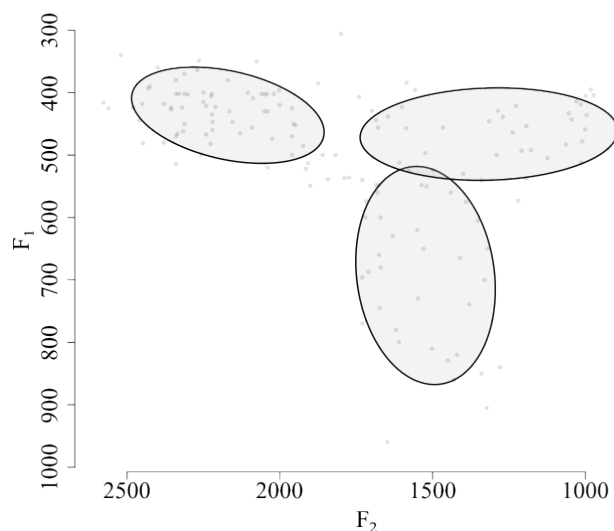


Figure 3.1: A mixture of two-dimensional Gaussian distributions. Multidimensional Gaussian distributions generalize the single-variable Gaussian distribution—a symmetrical probability distribution based on a “sum of squared error” computation, with a location parameter  $\mu$  setting the center, and a scale parameter  $\sigma^2$ , the variance, setting how quickly the probability falls off away from the center—to multiple dimensions. The center becomes a  $p$ -length vector, and the scale becomes a  $p$ -by- $p$  matrix, listing not only the variance on the  $p$  dimensions, but also their covariance (unscaled correlation). The set of points with total probability  $\pi$  is an ellipsoid which is aligned with the axes if the variables on the  $p$  dimensions are not correlated, and otherwise has some rotation in proportion to the degree of correlation. This mixture has three categories, all with the same covariance matrix. Each ellipse is a 66% confidence region: in any direction, the probability of the deviating farther than that from the center is the same as we go around the edge of the ellipse, and the total probability of all the points in the ellipse is 0.66.

from Gaussians; the model included a second layer implementing a lexicon where the tokens were assumed to be organized into sequences). Another clustering model that has been applied to phonetic category learning (the output of which can be interpreted as a non-Gaussian mixture model) is  $k$ -means (Hall & Smith 2006). Crucially, fitting all these models is done unsupervised: the learner is given a set of observations, and does not know any information about which tokens came from which categories; they must induce some

mixture model that fits the data well.

Fitting a mixture model, as we said, involves selecting category maps, which, given that those category maps need to be specified in some way or another, means finding some parameter values for each. The most common approach today to fitting a mixture model in a Bayesian way is to put a nonparametric Bayesian prior on the parameter values, which is a prior on the set of distributions over the parameter values; the most popular is the Dirichlet process prior. To understand the idea of choosing a prior—which is a probability distribution itself—on the set of probability distributions over parameter values, and how this relates to mixture models, consider a concrete example, a mixture of multivariate Gaussians. Think of it first as a set: if there are three categories, the set should have three pairs of parameter values (or, in our earlier version, three items, each associated with a single set of parameter values):  $\{\langle \underline{\mu}_1, \Sigma_1 \rangle, \langle \underline{\mu}_2, \Sigma_2 \rangle, \langle \underline{\mu}_3, \Sigma_3 \rangle\}$ . Each  $\Sigma$  gives the shape and orientation of a different set of ellipsoids, and each  $\underline{\mu}$  gives the center for them. Now, since we also need, as part of the process of generating an observation in the generative model, some probabilities for selecting each of the different categories (the bias we talked about—even if it is implicit:  $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ )—and since any probability distribution has got to be stated over some set anyway—we can just say that, when we are choosing a particular mixture model, we are actually choosing a probability distribution on parameter values—a distribution where it just so happens that all but a finite subset of the parameter values (in this case exactly three) have zero probability. When we are choosing a mixture model, we are looking for distribution over a set. Thus, to use Bayesian inference to decide on a mixture model, we need a prior distribution over probability distributions.

A Dirichlet process (Ferguson 1973) is such a distribution. When we draw a sample from a Dirichlet process, we get back a discrete distribution on the set of parameters, meaning that that distribution puts non-zero probabilities on a countable subset of the parameters; this is exactly the sort of distribution we want to characterize a mixture model, as it captures the idea that the “select a category” step is the selection of one of a collection of distinct items (there is not a continuous range of vowel phonemes in English, there are distinct equivalence classes). Although it might seem surprising that we have not said that the draws are distributions that put probability on finite sets of parameters, this is actually important, as it allows us to capture another fact, that of “held-out probability”: there is always an outside chance that we observe a new and totally unfamiliar category. This might not seem important or even correct for an adult speaker of a language, but it is crucially important in learning: the learner needs to be able to posit categories it may not have heard before while keeping high probability on its old posited category structure, and in order to do this coherently, mixtures with high probability after some learning has taken place need to assign some probability to the new category—otherwise, the mixture’s existing high probability could not influence its future posterior probability. This approach is not unreasonable for adult speech perception either: for an adult perceiver it might not be impossible to perceive a sound as belonging to a new category, even if the way speech perception works makes it unlikely. When we use a distribution drawn from a Dirichlet process prior as our way of selecting a category, we are said to have an infinite mixture model, because the model actually has infinitely many categories. (The word “process” applied to a probability distribution just means that draws are of infinite dimension.) Of course, the number of categories that will actually be assigned in a finite sample must be

finite!—and if this number were not generally substantially smaller than the sample size the Dirichlet process prior would not be very useful as a mixture prior—it is, although we can adjust how many categories we expect by changing the hyperparameters of the Dirichlet process.

To do this Bayesian inference—given some data and a prior on mixture models, find the parameters of a high-posterior mixture model—we would ideally be able to analytically solve for the maximum-posterior mixture. We cannot do this in this case, unfortunately, so we might then appeal to the idea of drawing a sample of many possible mixture models from the posterior distribution and then finding one that has very high posterior probability. We can almost do this—we use a type of Markov chain Monte Carlo, which is a way of drawing samples that have the posterior distribution as a limiting distribution. The MCMC technique we use is called Gibbs sampling, (Geman & Geman 1984), which walks over each of the different variables in a model, one by one, replacing the current value with a sample taken from the posterior distribution of that variable conditional on the current values of all the others. We will see that the crucial variables in our model are the parameter values, which need to be assigned, point by point, to the data, giving rise, implicitly, to the mixture. We Gibbs sample one parameter value assignment conditional on all the rest of them. The models will get more complicated, because we will add hyperparameters, but the essence of Gibbs sampling stays the same—given some initialization:

(73) Gibbs sampling (one step).

1.  $a \leftarrow$  a sample from the distribution of  $A|B = b, C = c, D = d$
2.  $b \leftarrow$  a sample from the distribution of  $B|A = a, C = c, D = d$
3.  $c \leftarrow$  a sample from the distribution of  $C|A = a, B = b, D = d$

4.  $d \leftarrow$  a sample from the distribution of  $D|A = a, B = b, C = c$

We do this for some number of steps, taking the  $\langle a, b, c, d \rangle$  tuple we are left with at the end of each iteration as a single sample. Usually we first run for a large number of iterations (“burn-in”), and use various rules of thumb to see whether the variables have actually converged to the posterior; after this we collect samples, throwing out a few in between each to correct for the fact that Gibbs samples are highly time-correlated (“lag”). If it is reasonable to calculate the value of the posterior for each sample we draw (which does not mean it was ever possible to maximize it analytically!) then we can pick the highest-posterior sample; we can also take an average where that is appropriate or possible (although it is not for our models).

We have discussed mixture models, and we have outlined how we will be estimating them from data. Since we have said that mixture models capture systems of phonetic categories, but we will be studying systems of phonetic category and allophony, it is worth reviewing some of the approaches that have been taken in the computational literature to learning allophony. There are basically two sorts. First, there are systems that directly learn phonological grammars of one kind or another. Second, there are systems that try and come up with criteria for detecting allophony, without giving a full account of how the grammar would actually be learned (actually, there is only one system of note, but it is used in a few different papers). We will put all of the supervised phonological grammar learning models aside, which is to say, models that need to be provided explicitly with the correct underlying lexical representation as part of the input for each observed form. This leaves basically two kinds of learners that deal with allophony: learners of phonotactic grammars,

and learners that follow the heuristic of Peperkamp, Le Calvez, Nadal & Dupoux 2006.

Phonotactic learning means learning sequencing restrictions on segments in a language. For example, in English, *blick* is a possible word, but *\*bnick* sounds very strange. That English speakers have this judgment reveals their knowledge of phonotactic restrictions, many of which are language-specific and learned. As was discussed briefly above, standard Optimality Theory enforces a constraint that underlying forms should be just like surface forms unless there is an alternation to support a discrepancy, so much of allophony is treated as constraints on surface representations, not as changes that take place in the grammatical mapping. These constraints will at any rate be found by surface phonotactic learners, such as the Hayes & Wilson 2008 learner and the phonotactic component of the word segmentation model of Blanchard & Heinz 2008. These learners adjust their expectations for phone sequences on the basis of observed sequences, so that, for example, in English, they would come to disprefer [#st<sup>h</sup>] sequences, with an aspirated stop out of position—strongly, presumably, because these will never occur in the pre-processed data that is usually presented to these learners (they could occur in a larger system subject to misperception).

The heuristic of Peperkamp et al. is aimed at looking in finer detail at these sequence distributions with the explicit goal of constructing a rule relating two segments. The idea is to make pairwise comparisons between segments on the basis of their immediate left and right contexts—for example, [t] will not occur after # in English. The full context-distributional profile of [t] reveals that it is in complementary distribution with [t<sup>h</sup>], and, as undergraduates learning phonology are instructed to do, a good learner is expected to take this as a cue to posit a rule relating the two in order to explain this fact. The Peperkamp et



al. heuristic is to quantify the discrepancy between the two context distributions in a way that is graded (using symmetrized KL-divergence), so if some illicit sequences actually do occur, the learner will still be able to detect that the distributional profiles of the allophonically related elements are quite different. If a pattern is truly allophonic, then it does seem clear that relatively high KL-divergence is expected. Noise or other patterns will obscure this to some degree. The effect of this heuristic should be predicted, in some form, by any system that evaluates predictions on the basis of proposed allophonic processes, because these hypotheses will themselves predict high KL-divergence.

Neither type of learning system is of immediate interest to us here. The proposal that I make in the next section takes allophony in a very different direction. In particular, it implies that there are no categorical surface representations (sequences of segments) that could be used to learn from the way these learners do. This will be discussed at greater length in Chapter 4 and defended empirically in Chapters 4 and 5. We move to that proposal now.

### 3.1.3 Phonetic transform hypothesis

We have discussed the basic architecture of the phonological system: there are lexical representations that are sequences of discrete-valued segments. There is a mapping that converts between these kinds of representations, on the one hand, and the kinds of representations that are used in receptive and productive systems, “phonetic representations,” a mapping called the phonological grammar. (Occasionally it is suggested that there are two different grammars, one for perception, and one for production, but it is usually un-

derstood that the two are not just mutually consistent but in some sense the same.) The phonological grammar gives rise to phonological alternations, which are changes to segments that the grammar makes depending on the context they appear in, like the Spanish [b]/[β] alternation, or the English alternate/alternation alternation, or the Hungarian vowel harmony alternation, discussed in Chapter 1. Almost universally, the grammars that phonologists provide treat these alternations as changes from one discrete-valued sequence to another discrete-valued sequence of segments.

On one end of the phonological grammar is the lexicon, and on the phonetic end of the phonological grammar the representations need to be consistent with “phonetic interpretation,” which is to say that they need to be the kinds of representations that the cognitive systems in perception and production work with. The receptive system needs to take perceptual representations and map them to lexical representations via the grammar, and the production system needs to take the output of the grammar and pronounce it. It has long been recognized, however, that phonetic cognition is not simply a matter of static “interpretation.” Rather, there also learned, context-dependent alternations that happen as part of a “phonetic grammar.” The reasoning is simply that there are some alternations that do not seem to be discrete-valued. The idea is that the phonetic interpretation component also has the ability to make changes based on the context, and these can be learned and language-specific (this does not do justice to the reasoning: we will discuss why something like this has to follow from this premise in Chapter 4). Claims of context-sensitive gradient phonetic changes under architectures that clearly also support discrete-valued alternations as well are to be found in Sledd 1966, Liberman & Pierrehumbert 1984, Port & O’Dell 1986, Cohn 1990.

For example, Liberman & Pierrehumbert 1984 contrast (1) the alternation between the two types of pronunciations of the English indefinite article, [ej]/[ə] versus [æn]/[ən], which depends (only) on whether the following word starts with a vowel or a consonant; and (2) the insertion of a short closure between [n] and [s] at the end of a syllable in American English, so that tense is pronounced something like [tnts]. For the first, they write that “the observed sound pattern is exactly what is expected if the phonological [category] representation contains an /n/ in one case and not in the other”; but, for the second, although one proposal might be a discrete-valued rule inserting the segment [t], yielding [tnts] at the surface representation, there are systematic differences between the “[t]” inserted in tense and underlying lexical [t] in the same environment, with the latter being systematically longer in duration. The crucial step in the reasoning is that this process does not reflect discrete categories, contrary to the usual understanding of what phonological rules do; but the first process does. The rule is gradient, thus it is understood to be part of a different, phonetic component of grammar. (If it seems unclear how this is distinguishable from an argument for a strict allophonic versus a neutralizing process, it is; see Chapter 4.)

To make this serious, we need a theory of context-sensitive phonetic operations. I will propose the outlines of one. I will assume a framework where the grammar specifies the operations directly (a “derivational” theory), rather than a framework, like Optimality Theory, where the grammar is represented as a set of constraints on the mapping, and then deduces what it needs to do for a given input. I believe this is an important step regardless of whether we think that phonetic grammars should be specified positively: we still need to understand what the “compiled-out” grammars that we could obtain can and cannot look like. Here is a strong hypothesis about the phonetic transforms that form the basic

operations of context-dependent phonetic grammar:

(74) Linear additive phonetic transform hypothesis (LPT). Phonetic transforms are additive and they are given by linear functions of the context.

There is a lot embedded in this statement. Let us go over it, point by point, starting from the idea that phonetic representations are “gradient,” in the sense that they can make much finer-grained distinctions than the discrete-valued categories we attribute to the lexicon.

**Phonetic representations** Call the set of all possible phonetic representations  $S$ .

**Addition of phonetic representations** There is an operation we call  $\oplus$  that applies to pairs of phonetic representations to give new ones.

We will put aside for the moment which of the properties associated with addition we will attribute to  $\oplus$ . Before saying what it means to be “linear,” we need to say what it means for a transformation to be a “function of the context.” What this means is that contexts have representations too (of course).

**Context representations** Call the set of all possible context representations  $A$ .

**Phonetic transformations** Call the set of all possible phonetic transformations  $\mathcal{T}$ . Transformations  $T \in \mathcal{T}$  are mappings  $T : A \rightarrow (V \rightarrow V)$ .

**Additivity** For any context  $\bar{a}$ , a transformation  $T$  gives rise to a unique phonetic representation  $t(T(\bar{a}))$  such that  $T(\bar{a})(\underline{r}) = \underline{r} + t(T(\bar{a}))$  for any phonetic representation  $\underline{r}$ .

Now we see how transformations can be functions of context. For them to be linear functions, we need to say something else about context representations.

**Addition of contexts** There is an operation we call  $\boxplus$  that applies to pairs of contexts to give new ones.

Context representations could either be phonetic representations or lexical-type (“phonological”) representations, or they could be qualitatively different from either, but they need to be recoverable from and convertible to the phonological representation output by the phonological grammar. If we make a firm statement here about their being exactly identical to one or the other, we would be placing constraints on what phonological or phonetic representations need to look like, and that is not the goal at present. However, the crucial assumption in everything that follows is that they at least can be phonological. We now say what it means for transformations to be linear in the context.

**Linearity** A transformation  $T$  is linear in  $A$  if  $t(T(\bar{a} \boxplus \bar{b})) = t(T(\bar{a})) \oplus t(T(\bar{b}))$ .

The usual is additional requirement added here is that we be able to “scalar multiply” the context representation and get a transformation that has a scalar multiple of the effect. Since we have not said anything about scalar multiplication existing for context representations, we will put this aside for the time being. Putting this together with additivity, we get the following.

**Linear additive transformations** For all contexts  $\bar{a}, \bar{b}$ ,  $T(\bar{a} \boxplus \bar{b})(\underline{r}) = \underline{r} \oplus (t(T(\bar{a})) \oplus t(T(\bar{b})))$ .

As we said, we do not want to say a lot about these two “addition” operations here and let this lead us to a lot of new conclusions about phonetic and phonological representations (up to now we have not even said that they are addition-like in any way). Rather, we want to take what we know about phonetic and phonological representations to constrain the

operations, so that any theoretical conclusions we draw are based primarily on this one crucial premise. Thus we would like to constrain both of the combinators based on some basic facts about what it would mean for a context to affect a phonetic representation. To preview: we will resort to using real space for both, out of convenience, so that we can continue to use mixture-of-Gaussian type models, because, once we appeal to the use of Gaussians, we are going to need to be working with real numbers; as a result we will just think of these two operators as regular addition. However, we would like for only the simulation arguments in this chapter to rest on that assumption, (out of technical necessity), and none of the theoretical arguments in the following two chapters.

Starting with  $\boxplus$ , the idea is that two different contexts can have effects that will combine in some way (according to  $\oplus$ ) whenever the context is really a combination (according to  $\boxplus$ ) of the two contexts. Without saying anything about how contexts are represented, the only circumstance in which two separate contexts should affect the same phonetic representation is if they are both present in the environment: an effect of preceding coronals and an effect of following uvulars should only be combined on a vowel when there is a coronal preceding it and a uvular following it. In our version, we will use discrete phonological representations, (for the most part), which are binary valued, and we will represent them as 0/1-valued vectors, one bit for each feature. That is why we will be able to interpret  $\boxplus$  as  $+$ : if we add a vector that is all zeroes except for the “preceding coronal” position to a vector that is all zeroes except for the “following uvular,” position, we will have a vector that is all zeroes except for the “preceding coronal” position and the “following uvular” position, interpreted as a conjunction of the propositions “has a preceding coronal” and “has a following uvular.” Real addition would also have a sensible

meaning if we valued the context dimensions as  $-1$  and  $+1$  (which we will do when we talk about feature-based learning models later in the chapter). Then, if we add  $+coronal$  to  $-coronal$ , we obtain  $0$  coronal, which we could interpret as removing the coronal feature from the representation, and thus, in this context, from the set of things that will affect the phonetic output.

Moving on to  $\oplus$ , the idea of “combining” phonetic representations needs to be taken in the context of what those phonetic representations give us. In terms of perception, we have been talking about the “phonetic maps” for individual segments. These are the recognition models that tell the listener which phonetic (in this case perceptual) tokens to expect, given that a particular category is being uttered. The representation we said was used for a phonetic map above was a Gaussian (which is probabilistic, so in fact it says more than which tokens to expect, but how much to expect it). This sets up a bit of a tension. When we say “phonetic tokens,” we really mean “phonetic representations” in some sense. However, when we talk about the “phonetic representation” of a category being learned, we mean that some Gaussian distribution is learned. These two things are in conflict. It is true that we can specify a Gaussian by giving only one “possible token,” standardly the mode (“center”) of the distribution. However, we cannot specify all Gaussians in this way; we can only specify Gaussians with some fixed size and shape. In general, we also need to specify these, via the covariance matrix, and, given the general acoustic shape of phonetic categories (Figure 3.1 above is typical), it seems like a good working assumption that the learner needs to do this too. The problem is that, to map out the category assignment of perceptual representations in a particular region, one needs to specify that region, which seems to be a different kind of information than just a single perceptual representation — at

the very least we need to provide more than one representation—so which type of representation is a “phonetic representation”? The information associated with a category about how to recognize it, or the information being recognized? The problem when we start to talk about combining phonetic representations is, therefore, what we are adding to what.

The usual Gaussian representation in terms of a location and scale gives us some helpful guidance as to how to resolve this. Suppose a Gaussian phonetic map of a single category really is what we want to call a “phonetic representation.” Then real-vector addition of two such representations moves the center of the phonetic map to the sum of the two centers, which may be outside the interpretable range for another category (think of adding two front vowels, both with a second formant of about 3000 Hz); it also changes the shape in a way that does not make very much sense: adding two covariance matrices gives a family of ellipses for which the principal axes have been summed both in their orientation (add the eigenvectors) and their squared lengths. Thus, adding a category to another category will not give a very good result. However, there is no phonological pattern that we wanted to treat as adding two categories anyway; rather, what we want to add are phonetic representations that may not be interpretable as categories. In the model presented here, the type of transform we will use just changes the location. If we maintain that the specification of the Gaussian is a “phonetic representation,” (which it must be in some sense), then to get the result that the location changes under vector addition, we need the transforms to be vectors containing some amount that will be added to the mean, and all zeroes where the covariance matrix would be affected. Thus the transforms we use have phonetic representations that could never really be categories, but we still consider them all phonetic representations, subject to (in this case) real-valued addition.



To sum up: we have put forward a hypothesis about what the basic context-dependent phonetic operations are, namely, that they are “additions” which combine in a way that preserves the “addition” structure of the contexts that generate them. We have used this to put some constraints on what the addition of contexts should look like; we have spelled out the fact that the structure of the problem requires phonetic representations that serve two different purposes, categories and transforms, to look somewhat different. But what is the value of the LPT if we made the “addition of contexts” operation to order and know nothing about what kind of “addition” might preserve this structure, except that we are required talk, unnaturally, about “adding” two types of phonetic representations that seem fundamentally different?

The key is in the notion that  $\boxplus$  could only be acting to combine contexts. We can continue our reasoning: if two different contexts are present, then we should be able to add them to get a context representation that states that both contexts are present (similarly for two absent contexts, the first present and the second absent, and so on). This is simply an accumulation of independent pieces of information, and, as such  $\boxplus$ , like addition, should be commutative and associative:

Associativity of  $\boxplus$   $(\bar{a} + \bar{b}) + \bar{c} = \bar{a} + (\bar{b} + \bar{c})$

Commutativity of  $\boxplus$   $\bar{a} + \bar{b} = \bar{b} + \bar{a}$

This has consequences:  $\oplus$  must also be associative and commutative, at least over the

phonetic representations that we get as the effects of transforms:

$$(75) \quad \underline{r} \oplus (t(T(\bar{a})) \oplus (t(T(\bar{b})) \oplus t(T(\bar{c})))) = \underline{r} \oplus ((t(T(\bar{a})) \oplus t(T(\bar{b}))) \oplus t(T(\bar{c})))$$

$$\underline{r} \oplus (t(T(\bar{a})) \oplus t(T(\bar{b}))) = \underline{r} \oplus (t(T(\bar{b})) \oplus t(T(\bar{a})))$$

This is different from the types of changes we see in phonological grammars: changes in particular environments from one segment to another, or additions/deletions. These are associative (if we have three processes and want to convert the action of two of them a single process, it does not matter which two we pick), but they are not commutative: changing  $A \rightarrow B$  and then  $B \rightarrow C$  does not give the same result as changing  $B \rightarrow C$  and then  $A \rightarrow B$  (see Chapter 1). This has nothing to do with whether phonological grammars are monostratal or derivational; it is just a fact about the kinds of changes we see on strings, namely, that one segment in one environment will change in one way, while another will change in a different way, and this breaks commutativity of composition in general. If it never happened that grammars needed to be stated as compositions of multiple operations, and we therefore never had to face this fact, it would still be a fact, and the fact would therefore remain that, under this view, phonetic grammars are fundamentally different from phonological grammars.

Now, there are some reasons we might think that  $\boxplus$  might not be commutative: suppose that contextual information (assuming that it is discrete) is organized into feature hierarchies, as some phonologists have suggested. Then it might not make sense to “add” a feature to contextual information without first adding its parent in the tree. Then the effect on the operation of  $\oplus$  would be different, but still predictable, namely, that the order of

composition would necessarily track the hierarchical organization of features. In any case, LPT has the effect of constraining the operation of phonetic transformation according to the means by which contextual information can be combined.

This will be cashed out in the model presented in the next section as follows: since, under our assumptions of convenience about how to interpret all of this as real-valued Gaussian phonetic maps, we have decided that we are only going to allow phonetic transforms to affect the location of the Gaussian (so, they always add zero to the covariance matrix), we can state a phonetic category as a Gaussian linear model: the context is a vector for us, so separate all the different dimensions out into components  $x_1, x_2$ , and so on, up to  $x_{h-1}$ . Then the transformed location of a phonetic category map that starts at  $\underline{a}_0$ , in context, is:

$$(76) \quad T(\underline{a}_0) = \underline{a}_0 + x_1 \underline{a}_1 + \cdots + x_{h-1} \underline{a}_{h-1}$$

If  $\bar{x}$  is an augmented context vector,  $\langle 1, x_1, \dots, x_{h-1} \rangle$ , then we can write this as a matrix multiplication:

$$(77) \quad T(\underline{a}_0) = A^T \bar{x}$$

—where  $A$  is an  $h \times p$  matrix ( $p$  being the number of dimensions of the phonetic location vector) where the first row is  $\underline{a}_0$ , the second is  $\underline{a}_1$ , and so on. This matrix will represent the category  $A$ , along with all the transformations that can apply to it.

I finish this section with the following conjecture, which will be explored (and ex-

plained) in more detail in Chapters 4 and 5: allophony is phonetic grammar. This says that the output of the phonological grammar does not contain any of the information that is phonetically present but not contrastive lexically, like the epenthetic stop found after the [n] in English tense. It does contain the result of other processes, like the [ɪ] before the -ion in English alternation. The details of what exactly should count as an “allophone”—apart from saying that it is a pronunciation that can only ever occur as the result of a particular context-dependent process, and not lexically—will be left for Chapter 4.

The conjecture is not new: “It has been our experience,” write Liberman and Pierrehumbert, “that cases of ‘allophonic variation’ often turn out to have properties like those of [the English tense phenomenon]. This leads us to suspect that a correct division of labor between phonological representation and phonetic implementation will leave the output of the phonology rather more abstract than it is usually assumed to be” (228–229); Kiparsky 1985 pointed to this suggestion as being of potential interest to phonologists. However, as far as I know, this dissertation is the first place the consequences for phonological theory or for phonological acquisition have been worked out.

## 3.2 A computational model: Dillon, Dunbar and Idsardi (2013)

### 3.2.1 Mixture of linear models

Mixture models over some set of observables  $Y$  can all be divided into three parts:

- (78)
- |            |                                       |
|------------|---------------------------------------|
| $\Theta$   | a set of possible parameter values    |
| $P$        | a discrete distribution over $\Theta$ |
| $F \theta$ | a distribution over $Y$               |

The distribution  $F|\theta$  is a “category model”: each different value of  $\theta$  gives rise to a different set of expectations about the observables, and we call such a set of expectations a category model. The distribution  $F|\theta$  might be specified as a Gaussian distribution on  $Y$  with location parameter  $\theta$  if  $\Theta$  were real numbers that could act as the location; or as a Gaussian distribution on  $X$  with parameters  $\theta = \langle \mu, \sigma \rangle$  if  $\Theta$  were made up of pairs of real numbers with positive real numbers that could act as the location and scale respectively; or as some other distribution on  $Y$  that somehow depends on  $\theta$ . The point is that different possible values of  $\theta$  specify different categories: for each, there is a different distribution  $F|\theta$  over the observables, which in our case are percepts. The distribution  $P$  gives our expectations about which of these categories will be realized, and it is discrete, which means that there are countably many categories. In all the cases we care about, the number of categories is really finite, but, as discussed above, moving to the countable case allows us to handle what it means for a percept to belong to a “previously unseen category” neatly.

Under the phonetic transform hypothesis, we assume that this phonetic model for a category—a single segment in the perceptual inventory—changes depending on the “environment,” which means some temporal window of context around the segment; furthermore, we assume that this change takes the form of a simple vector addition to the

location of the category which is a linear combination of all the pieces of information available about the context (“linear transform hypothesis”).

In this model, we assume that a system of categories is actually a system of composite objects: the mixture model for such a system is a mixture of linear models. From now on, we will illustrate these as in the rightmost picture in Figure 3.2. Two conventional mixture models are shown in the other two pictures, to draw the contrast. Each category in the conventional mixture models is drawn as an ellipse (in this case, the ellipses delimit the fifty percent confidence regions of two-dimensional Gaussians). However, in the mixture of linear models, there is a shift in the category which depends on the context. In this case, we have shown a model sensitive to one environment, and as the context information was assumed to be a simple indicator (zero or one) in this particular system, we have illustrated the context-dependence by drawing two ellipses, one solid and one dotted, for each category. The solid lines show the categories in the environment represented as  $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$  and the dotted lines show the categories in the environment coded as  $\begin{bmatrix} 1 & 1 \end{bmatrix}^T$ . See above for more details.

The explicit representation of a single category in a mixture model is a grammar for that category, which we hypothesize to share structure with the brain (in this case, the perceptual system) under some reasonably strong homomorphism. It specifies the category. In a mixture of linear models, it is a particular parameter matrix  $A$ , which contains the intercepts and the effects of different contextual features. The locations of the ellipses shown in the third panel in Figure 3.2, on the other hand, are not explicitly represented, except for the intercept. Rather, these ellipses are epiphenomenal, in the sense that they

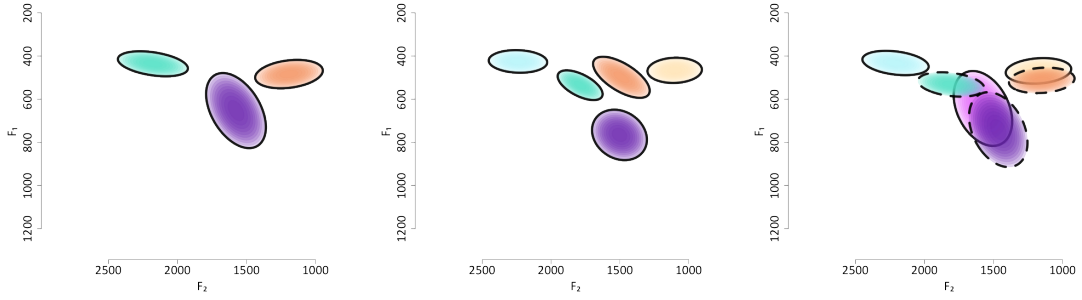


Figure 3.2: Illustrations of two different conventional mixture models (left and center) and of a mixture of linear models (right). In the conventional mixture model, each category corresponds to a single distribution over the observables. In the mixture of linear models, each category is complex, and gives rise to a family of possible distributions, with different locations. In particular, the location is a linear combination of the contextual information. In this case, only two ellipses are shown, because this model shows a simple case where there is one piece of contextual information, which is a single bit.

constitute derived structure. We make no claim about their cognitive reality, and the strong claim is that they have none.

If we start with a vector Gaussian category model on  $p$  dimensions, assuming a mixture of linear models means that we move from  $F|\theta$  being a distribution specified by this density function—

$$(79) \quad f(\underline{y}|\underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \right\}$$

—to this, where  $\bar{x}$  is again the vector coding all the information about the context, where the first element is always 1, “member of the category in question”:

$$(80) \quad f(\underline{y}|A, \Sigma, \underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\underline{y} - A^T \bar{x})^T \Sigma^{-1} (\underline{y} - A^T \bar{x}) \right\}$$

Given this, a Bayesian mixture model under a Dirichlet process prior is as follows:

$$\begin{aligned}
 (81) \quad & G \sim DP(\alpha G_0) \\
 & A_i, \Sigma_i \sim G \\
 & \underline{y}_i | A_i, \Sigma_i \sim \mathcal{N}(A_i^T \bar{x}_i, \Sigma_i)
 \end{aligned}$$

This turns out to be a special case of the dependent Dirichlet process prior (MacEachern 1999). The conjugate prior on  $A, \Sigma$  is the inverse Wishart distribution compounded with the matrix normal distribution (Dawid 1981). This is a generalization of the conjugate prior in the case where the mean follows a multivariate (rather than a matrix) normal distribution, the normal-scaled-inverse-Wishart prior. In that prior, the distribution on the mean has covariance  $\Sigma$  (the data covariance) except by a constant  $\omega$ : if we use a wholly different covariance matrix we break conjugacy. Here we replace  $\omega$  with a full covariance matrix  $\Omega$ , which specifies the variance on each row of  $A$  as well as the covariances between rows. This is the matrix normal density, where  $h$  is total length of the context vector  $\bar{x}$ :

$$(82) \quad f(A|A_0, \Sigma, \Omega) = \frac{1}{(2\pi)^{\frac{hp}{2}} |\Sigma|^{\frac{h}{2}} |\Omega|^{\frac{p}{2}}} \exp \left\{ -\frac{1}{2} \text{tr} \left\{ \Sigma^{-1} [A - A_0]^T \Omega^{-1} [A - A_0] \right\} \right\}$$

When  $A$  follows a matrix normal distribution we will write  $A \sim \mathcal{N}(A|A_0, \Sigma, \Omega)$ . Adding



several hyperparameters, we thus obtain the following model:

$$\begin{aligned}
(83) \quad & \alpha \sim \text{Gam}(\alpha|a, b) \\
& A_0 \sim \mathcal{N}(A_0|M_0, S_0, \mathbb{I}_h) \\
& \Omega \sim \mathcal{IW}(\Omega|\Phi, \lambda) \\
& G_0 = \mathcal{N}(A|A_0, \Sigma, \Omega) \times \mathcal{IW}(\Sigma|\Psi, \kappa) \\
& G \sim DP(\alpha G_0) \\
& A_i, \Sigma_i \sim G \\
& y_i|A_i, \Sigma_i \sim \mathcal{N}(A_i^T \bar{x}_i, \Sigma_i)
\end{aligned}$$

One Gibbs sampler (the “Chinese restaurant process” construction) works like this: we first add an arbitrary index  $z_i = z_j \Leftrightarrow \langle A_i, \Sigma_i \rangle = \langle A_j, \Sigma_j \rangle$  drawn from some other set. We then sample as follows, integrating out  $G$ :<sup>1</sup>

---

<sup>1</sup>Note that the normalizing constant for this distribution over possible values of  $z_i$  is not the one that makes the integral in this second clause come to one but rather one that makes the probabilities over all the logically possible  $z_i$  values come to one, which is equal to the sum of all the different probabilities for existing  $z_j$ , as given in the first clause, plus the probability of a different value, given in the second clause. The integral in the second clause (the “likelihood integral”) does not come out to one, but since we know the form of each of the three terms, and since we know the form of a function that does integrate to one and differs from the integral only by a constant (namely,  $f(A, \Sigma|y_i, \bar{x}_i)$ , a normal inverse Wishart density again, by conjugacy), the integral can easily be computed by pulling out all the constant terms to yield  $w \int s$ , doing the same with the posterior density to get  $z \int s$ , and then noting that  $wz^{-1}z \int s = wz^{-1}$ . This is how to obtain the ratio given in the text.

(84) Algorithm 1.

1. For  $i$  in  $1, \dots, N$  (where  $N$  is the total number of data points):

(a)  $z_i$  is resampled from the distribution:

- $P(z_i = z) \propto N_z \cdot f_{\mathcal{N}}(y_i | A_i^T \bar{x}_i, \Sigma_i)$ , where  $N_z$  is the number of  $z_j = z$  and  $f_{\mathcal{N}}$  is the normal density, when  $z = z_j$  for some  $j \neq i$
- $P(z_i = z) \propto \alpha \cdot \int_{A, \Sigma} f_N(y_i | A^T \bar{x}_i, \Sigma) \cdot f_{\mathcal{N}}(A | A_0, \Sigma, \Omega) \cdot f_{\mathcal{IW}}(\Sigma | \Psi, \kappa) dA, d\Sigma$ , where  $f_{\mathcal{IW}}$  is the inverse Wishart density, when  $z$  is different from all  $z_{j, j \neq i}$ .

· The integral works out to  $\Gamma_p\left(\frac{\kappa'}{2}\right) |\Psi|^{\frac{\kappa'}{2}} / \pi^{\frac{p}{2}} \Gamma_p\left(\frac{\kappa}{2}\right) |\mathbb{I}_h + \Omega \bar{x}_i \bar{x}_i^T|^{\frac{p}{2}} |\Psi'|^{\frac{\kappa'}{2}}$ , where

$$\Psi' = \Psi + A_0^T \Omega^{-1} A_0 + y_i y_i^T - [\bar{x}_i y_i^T + \Omega^{-1} A_0]^T [\Omega^{-1} + \bar{x}_i \bar{x}_i^T]^{-1} [\bar{x}_i y_i^T + \Omega^{-1} A_0] \text{ and } \kappa' = \kappa + 1.$$

(b) If  $z_i$  is distinct from all  $z_{j, j \neq i}$ ,  $\langle A_i, \Sigma_i \rangle$  is resampled from  $\text{Pr}_{A, \Sigma}[\cdot | y_i]$ , that is, the normal-inverse Wishart prior on  $A, \Sigma$ , multiplied by  $\frac{f_{\mathcal{N}}(y_i | A^T \bar{x}_i, \Sigma)}{\int_{A, \Sigma} f_{\mathcal{N}}(y_i | A^T \bar{x}_i, \Sigma) \cdot f_{\mathcal{N}}(A | A_0, \Sigma, \Omega) \cdot f_{\mathcal{IW}}(\Sigma | \Psi, \kappa) dA, d\Sigma}$ . This is another normal-inverse Wishart distribution, with parameters:

$$A'_0 = [\Omega^{-1} + \bar{x}_i \bar{x}_i^T]^{-1} [\bar{x}_i y_i^T + \Omega^{-1} A_0], \Omega' = [\Omega^{-1} + \bar{x}_i \bar{x}_i^T]^{-1}, \Psi', \kappa' \text{ as above}$$

2. For each distinct  $z$ , let all  $\Sigma_i$  such that  $z_i = z$  be a sample  $\Sigma_z$  from an inverse Wishart distribution with parameters:

$$\begin{aligned} \Psi'_z &= \Psi + Y_z Y_z^T + A_0^T \Omega^{-1} A_0 - [X_z Y_z^T + \Omega^{-1} A_0]^T [\Omega^{-1} + X_z X_z^T]^{-1} [X_z Y_z^T + \Omega^{-1} A_0] \\ \kappa'_z &= \kappa + N_z \end{aligned}$$

where  $N_z$  is the number of  $z_i = z$ ,  $X_z$  is an  $h \times N_z$  matrix consisting of all  $\bar{x}_i$  for which  $z_i = z$ ,  $Y_z$  is the  $p \times N_z$  matrix consisting of all  $y_i$  for which  $z_i = z$  corresponding to  $X_z$

3. For each distinct  $z$ , let all  $A_i$  such that  $z_i = z$  be a sample  $A_z$  from a normal distribution with parameters:

$$\begin{aligned} A'_{0,z} &= [\Omega^{-1} + X_z X_z^T]^{-1} [X_z Y_z^T + \Omega^{-1} A_0] \\ \Sigma'_z &= \text{the } \Sigma_i \text{ shared by the } z_i = z \\ \Omega'_z &= [\Omega^{-1} + X_z X_z^T]^{-1} \end{aligned}$$

4. Sample  $\Omega$  from an inverse Wishart distribution with parameters:

$$\Phi' = \Phi + \sum_{z \in Z} (A_z - A_0)^T \Sigma^{-1} (A_z - A_0), \quad \lambda' = \lambda + |Z|p$$

where  $Z$  is the set of all distinct  $z_i$

5. Sample  $\text{vec}(A_0)$ , the  $hp$ -dimensional column vector obtained by concatenating the columns of  $A_0$ , from a normal distribution with parameters:

$$\begin{aligned} \text{location} &= \left[ [S_0 \otimes \mathbb{I}_h]^{-1} + \left[ \left( \sum_{z \in Z} \Sigma_z^{-1} \right)^{-1} \otimes \Omega \right]^{-1} \right]^{-1} \left[ [S_0 \otimes \mathbb{I}_h]^{-1} \text{vec}(M_0) + \sum_{z \in Z} [\Sigma_z \otimes \Omega]^{-1} \text{vec}(A_k) \right] \\ \text{scale} &= \left[ [S_0 \otimes \mathbb{I}_h]^{-1} + \left[ \left( \sum_{z \in Z} \Sigma_z^{-1} \right)^{-1} \otimes \Omega \right]^{-1} \right]^{-1} \end{aligned}$$

6. Sample a value  $x$  from a beta distribution with parameters  $\alpha + 1, N$ . Then sample  $\alpha$  from a gamma distribution:

- with parameters  $a + |Z|, b - \log x$ , with probability proportional to  $(a + |Z| - 1)$
- with parameters  $a + |Z| - 1, b - \log x$ , with probability proportional to  $N(b - \log x)$

The addition of the index variable allows us to modify the parameter values without changing the clustering (meaning the “association” of  $y_i$  with particular  $\langle A_i, \Sigma_i \rangle$ ); this allows us to capitalize on the fact that the clustering is often largely correct even when the

parameters are slightly off, which makes the sampler more efficient (Bush & MacEachern 1996).<sup>2</sup>

### 3.2.2 Summary of Inuktitut experiments

In Dillon, Dunbar & Idsardi 2013, we reported a number of statistical learning experiments using data from Inuktitut, an Eskimo-Aleut language, one of the three official

---

<sup>2</sup>It may be convenient to collapse out the parameter values entirely, and sample only the indices (MacEachern 1994). As this makes the conditional distributions on the hyperparameters of  $G_0$ ,  $\Omega$  and  $A_0$ , non-conjugate, however, a Gibbs sampler becomes not only computationally intensive but also inefficient (since the sampling step for the hyperparameters must then itself use MCMC or rejection sampling). With fixed hyperparameters, a sampler would look like this: (85) Algorithm.

1. For  $i$  in  $1, \dots, N$  (where  $N$  is the total number of data points),  $z_i$  is resampled from the distribution

$$P(z_i = z, \text{ for } z = z_j \text{ for some } j \neq i) \propto N_z \cdot \int_{A, \Sigma} f_N(\underline{y}_i | A^T \bar{x}_i, \Sigma) \cdot f(A, \Sigma | Y_z, X_z, A_0, \Omega) dA, \Sigma$$

where  $f(A, \Sigma | Y_z, X_z, A_0, \Omega)$  is the posterior normal Inverse Wishart density. The resulting integral works out to  $\frac{\Gamma_p(\frac{\kappa''}{2}) |\Omega''|^{\frac{p}{2}} |\Psi^*|^{\frac{\kappa^*}{2}}}{\pi^{\frac{p}{2}} \Gamma_p(\frac{\kappa^*}{2}) |\Omega^*|^{\frac{p}{2}} |\Psi''|^{\frac{\kappa''}{2}}}$ , where

$$\begin{aligned} \Psi^* &= [\Psi + Y_z Y_z^T] + A_0^T \Omega^{-1} A_0 \\ &- [X_z Y_z^T + \Omega^{-1} A_0]^T [\Omega^{-1} + X_z X_z^T]^{-1} [X_z Y_z^T + \Omega^{-1} A_0] \end{aligned}$$

$$\begin{aligned} \Psi'' &= [\Psi + Y_z Y_z^T + \bar{y}_i \bar{y}_i^T] + A_0^T \Omega^{-1} A_0 \\ &- [\bar{x}_i \bar{y}_i^T + X_z Y_z^T + \Omega^{-1} A_0]^T [\Omega^{-1} + X_z X_z^T + \bar{x}_i \bar{x}_i^T]^{-1} [\bar{x}_i \bar{y}_i^T + X_z Y_z^T + \Omega^{-1} A_0] \end{aligned}$$

$$\Omega^* = [\Omega^{-1} + X_z X_z^T]^{-1}, \quad \Omega'' = [\Omega^{-1} + X_z X_z^T + \bar{x}_i \bar{x}_i^T]^{-1}$$

$$\kappa^* = \kappa + N_z, \quad \kappa'' = \kappa + N_z + 1$$

$$P(z_i = z, \text{ for } z \text{ different from all } z_{j, j \neq i}) \propto \alpha \cdot \int_{A, \Sigma} f_N(\underline{y}_i | A^T \bar{x}_i, \Sigma) \cdot f_{\mathcal{N}}(A | A_0, \Sigma, \Omega) \cdot f_{\mathcal{IW}}(\Sigma | \Psi, \kappa) dA, \Sigma \text{ as before.}$$

2. Sample a value  $x$  from a beta distribution with parameters  $\alpha + 1, N$ . Then sample  $\alpha$  from a gamma distribution:

- (a) with parameters  $a + |Z|, b - \log x$ , with probability proportional to  $(a + |Z| - 1)$
- (b) with parameters  $a + |Z| - 1, b - \log x$ , with probability proportional to  $N(b - \log x)$

Given a sample consisting of a sequence of indices, (say, the highest-posterior sample),  $A, \Sigma$  for a given cluster can be maximized analytically under the normal inverse Wishart posterior. Sampling is generally somewhat less computationally intensive than updating the values of the various factors in the integral, but might be expected to be less efficient; however, my own experience indicates that it is difficult to find good fixed values for  $\Omega$ , and so I will use the algorithm in the text, and variants on it, throughout.

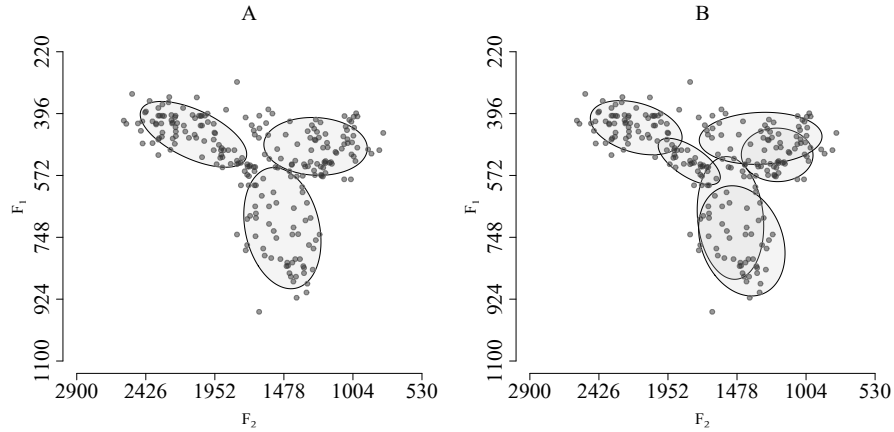


Figure 3.3: Vowel tokens from Inuktitut (see below for a description of the corpus). The dimensions are the second and first formant, which correspond closely to vowel backness and height. The ellipses in panel A are 66% confidence regions for multivariate Gaussians estimated by maximum likelihood for just the /i/ tokens, just the /a/ tokens, and just the /u/ tokens. In panel B, the ellipses are for Gaussians estimated separately for the tokens of [i], [e], (/i/ tokens out of, and in, the uvular context), [a], [ʔ], [u], and [o].

languages of Nunavut territory in Canada. Inuktitut has three lexically contrastive vowels, /i/, /u/, and /a/. An allophonic process of Inuktitut retracts the tongue root during the pronunciation of vowels when preceded or followed by a uvular consonant (Dorais 1986; Denis & Pollard 2008). For the sake of having some notation, I will say that, in this environment, /i/, /u/, and /a/ are pronounced as [e], [o], and [ʔ], respectively; see Figure 3.3.

The process is morphologically productive in the sense discussed in Chapter 1: the word for “pen” is *titirauti*, /titiauti/, which is pronounced in citation form as [titeauti], with a final [i]; embedded before another morpheme starting with a [t], it keeps this pronunciation, as in “your pens,” *titirautitit*, [titeautitit]; but before [q], as in “do you have a pen?,” *titirautiqagtunga*, [titeauteqqtuŋa].

We reported the results of three statistical experiments using data from Inuktitut: Experiment 1 fit a mixture of Gaussians, by sampling using Algorithm 1 followed by the

selection of a high-posterior sample, to data from Inuktitut (with no predictors); Experiment 2 fit the model to the same data, but with the uvular-environment tokens corrected for retraction: the numerical difference between the corpus average  $\langle F_1, F_2 \rangle$  in the retraction environment, and out of the retraction environment, was removed from all of the tokens appearing in the retraction environment, simulating a perceptual correction; finally, Experiment 3 used Algorithm 1 on the original, uncorrected data, with a single predictor, coded as 1 if a token appeared in a uvular environment, and 0 otherwise.

In all three experiments, the solutions reliably had three categories corresponding well to the phonemes of Inuktitut. However, we argued that the models in Experiments 2 and 3, which suggested versions of the phonetic transform hypothesis, were better because they captured the lawful relations between certain types of vowel tokens (those belonging to a particular vowel, in, and out of, the uvular environment). Although learners do gain receptive knowledge of this relation, it is not obvious how to recover this relation from a three-category mixture of Gaussians which obliterates the distinctions between allophonic categories. By increasing the number of data points to obtain a larger number of categories, (see Antoniak 1974), we then also showed that the five- and six-category solutions found by a mixture of Gaussians, although they resembled the allophones of the Inuktitut vowels somewhat, (with or without the subtle [a]/[ ] distinction), were too different from the actual allophones to remain in the contextual pattern: the assignment of vowel tokens to learned categories was such that the tokens that were assigned to categories in similar locations to [e,,o] were no longer reliably points that actually occurred in the uvular environment, and similarly for [i,a,u]. Experiment 3, on the other hand, showed a model which learned the phonetic content of the retraction rule (via the regression ma-

trices A), while the output of the mixture of Gaussians in Experiment 1 would evidently not be very useful for learning about the retraction rule, since the distributions are too inaccurate. This poor alignment with clear statistical clusters corresponding to phones is somewhat different from the results of other mixture of Gaussians studies of phonetic category learning, (Vallabha, McClelland, Pons, Werker & Amano 2007, Feldman, Griffiths & Morgan 2009), but those studies' data were collections of points sampled from Gaussians estimated to real-life phoneme category data; this makes the data fit the assumptions of the model, which makes it easier for it to learn. We used the raw data directly: 239 measurements of  $\langle F_1, F_2 \rangle$  at the steady state taken in Praat in a study on Inuktitut phonetics, (Denis & Pollard 2008), in single-word elicitations from one female Inuktitut speaker from Kinggait. (Upsampling to increase the number of data points was done nonparametrically using a two-dimensional kernel density estimate, according to natural mixing proportions obtained from the Nunavut Hansard corpus: Martin, Johnson, Farley & MacLachlan 2003.)

A summary of how the models estimated in Experiments 1–3 compare to the true classification of vowel tokens in the corpus is given in Table 3.1. The classification performance showed statistically significant improvement for the three-category solutions with respect to Experiment 1 in both Experiments 2 and 3, and for all the models (with whatever number of categories) in Experiment 3. This was one reason we gave for thinking that a phonetic transform model of allophony better supported learning perceptual categories than a model where allophony is categorical: the classification models we obtained when allophony was learned as a phonetic transform were slightly better than those obtained when it was not, and these models are thus presumably more like the native speaker's perceptual model. We attributed this improvement to the fact that these models can handle

the fact that lexical categories have multiple contextually determined statistical modes.

The other argument we gave in support of the phonetic transform hypothesis was that the six-category mixture of Gaussian solutions found in Experiment 1 were not well enough aligned with the allophones to preserve the quasi-complementary distribution between allophones: we computed symmetrized KL divergence scores between categories over the probability of being/not being in a uvular environment to assess the degree of complementarity in their distributions (following Peperkamp, Le Calvez, Nadal & Dupoux 2006). The KL divergence scores showed inconsistencies due to incorrect classifications that meant that pairs that ought to have had relatively low KL-divergence (low degree of complementarity, thus less surface evidence for allophony) had higher KL-divergence, and pairs that ought to have had relatively high KL-divergence (high degree of complementarity, thus more surface evidence for allophony) had lower KL-divergence. These statistics are summarized for the estimated six-category model in Table 3.2. Graphs of representative fitted models are shown in Figure 3.4.

Supervised baseline									
	$F$	Precision	Recall						
1000 data points	0.84	0.83	0.85						
vs allophones	0.64	0.66	0.63						
12000 data points	0.79	0.79	0.79						
vs allophones	0.69	0.64	0.76						
Raw (239 points)	0.78	0.78	0.78						
vs allophones	0.68	0.63	0.74						
Experiment 1									
	$F$	Precision	Recall	$K = 1$	2	3	4	5	6
1000 data points	0.70 (+.02)	0.66 (+.01)	0.74 (+.03)	0	0.125	0.625	0.125	0	0
vs allophones	0.60 (+.01)	0.50 (+.01)	0.76 (+.02)						
12000 data points	0.65 (+.01)	0.68 (+.01)	0.63 (+.02)	0	0.1	0.5	0.1	0.2	0.1
vs allophones	0.58 (+.02)	0.53 (+.01)	0.63 (+.02)						
Raw (239 points)	0.65 (−.03)	0.59 (−.03)	0.76 (−.03)	0.1	0.2	0.7	0	0	0
vs allophones	0.47 (−.04)	0.34 (−.01)	0.81 (−.02)						
Experiment 2									
	$F$	Precision	Recall	$K = 1$	2	3	4	5	6
1000 data points	0.73 (+.02)	0.67 (+.02)	0.82 (+.02)	0	0.4	0.5	0.1	0	0
12000 data points	0.74 (+.02)	0.72 (+.02)	0.76 (+.02)	0	0	0.875	0.125	0	0
Raw (239 points)	0.63 (−.01)	0.49 (−.02)	0.88 (−.01)	1.0	0	0	0	0	0
Experiment 3									
	$F$	Precision	Recall	$K = 1$	2	3	4	5	6
1000 data points	0.75 (+.02)	0.71 (+.02)	0.80 (+.02)	0	0.111	0.889	0	0	0
12000 data points	0.69 (+.02)	0.65 (+.02)	0.76 (+.02)	0.143	0	0.571	0.286	0	0
Raw (239 points)	0.69 (−.01)	0.64 (−.00)	0.79 (−.01)	0.125	0.125	0.75	0	0	0

Table 3.1: Summary of model evaluations from Dillon, Dunbar & Idsardi 2013, showing pairwise agreement between models and true classifications as  $F$ , precision, and recall scores (pairwise agreement: for each pair of data points, agreement or failure to agree on whether they belong to a single category); and the distribution of number of categories for the MAP estimate across ten chains, run with complementary 10% held out test sets. See Dillon, Dunbar & Idsardi 2013 for details. Supervised baseline: Gaussian estimated on the points from a given category; pairwise agreement vs three-way phoneme classification except where noted.



	[i]	[e]	[u]	[o]	[a]	[]
[i]	0	0.810	<u>0.033</u>	0.321	<u>0.330</u>	0.846
[e]	–	0	<u>0.478</u>	<u>0.098</u>	<u>0.093</u>	0.000
[u]	–	–	0	<u>0.138</u>	<u>0.143</u>	<u>0.504</u>
[o]	–	–	–	0	<u>0.000</u>	0.109
[a]	–	–	–	–	0	<u>0.104</u>
[]	–	–	–	–	–	0

Table 3.2: Complementarity of estimated categories’ context distributions (i.e.,  $P(\text{token is in uvular environment})$ , measured by symmetrized KL-divergence. Values computed over a six-category mixture classification found in Experiment 1; categories labelled by visual inspection to map to the closest actual allophone. The bold values are pairs that are actually allophones, and thus should have relatively high scores; the underlined values are pairs that should not be labelled as allophones, because they are members of the same retracted/non-retracted class. All of the underlined scores would ideally be substantially lower than all of the bolded scores, but this is not the case: the classification under this model obscures the complementarity of the distributions.

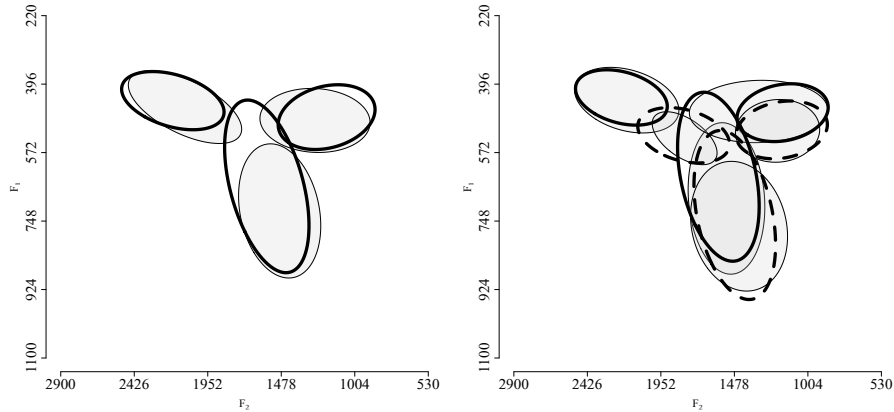


Figure 3.4: Example fitted models from Experiments 1 (left) and 3 (right) of Dillon, Dunbar & Idsardi 2013. The shaded ellipses are the supervised baseline (two different baselines, phonemic, and allophonic, are used to highlight the relevant underlying statistical patterns).

### 3.3 Selecting transform environments

#### 3.3.1 Mixture of linear models with variable selection

In a real-life situation, the learner does not know which of the large set of possible predictor values have an effect on pronunciation, and which do not: the learner does not know what environments trigger allophony. Since there is an identity  $\underline{0}$  in the hypothesis space for any given phonetic transform, (at least in our real-valued formulation), we might think that we do not need to know anything beyond what we have already said: if there is no effect, then the best estimate of the phonetic effect ought to be around zero anyway. However, from a practical perspective, it seems likely that the computational cost of not having to search for all the logically if some of them were obviously negligible would be improved; this would certainly be useful for our Gibbs sampler if there were large  $A$  matrices to be sampled. More importantly, it would also trigger a Bayesian Occam's Razor effect—the learner would be doing a model evaluation exactly like the toy example of the lexical decision task in Chapter 2. This would be true even if the distinction between setting phonetic transforms to zero and removing them entirely had no impact on any aspect of perception or production—that is, if this aspect of our formulation of the grammar was not part of the shared structure as regards the function the grammar computes. It would nevertheless be a part of the structure shared with learning, because it would impact the evaluation measure, if not the actual computational cost of making the evaluation. We would then be making a very clear (not necessarily easy to test) empirical prediction about learners' preferences. See Chapter 2.

The standard way to select different subsets of some available variables in a Bayesian way (the variable selection problem) is what is called a spike and slab model (Brown, Vanucci & Fearn 1998). This addition to the model simply attaches an indicator  $\gamma_j$  to each of the predictor variables  $x_j$ . If all  $\gamma_j = 1$ , then the model is exactly as before. However, if some  $\gamma_j = 0$ , then the model fixes the effect of predictor  $x_j$  to be exactly zero. The probabilistic interpretation of this is as follows: if  $\gamma_j = 1$ , then the conditional distribution of the  $j^{\text{th}}$  row of  $A$  is just as before—we say that this row is drawn “from the slab.” If  $\gamma_j = 0$ , however, then the conditional distribution of that row is a degenerate distribution with all of its mass at zero, the “spike.”<sup>3</sup> It is easy to show that for a Gaussian distribution (like the distribution on the  $A$  matrix), the distribution of any of the components conditional on the others is still Gaussian. Since degenerate distributions at a particular point can be seen as Gaussians with zero variance centered at that point, we obtain, conditional on  $\gamma$ , a degenerate matrix Gaussian distribution where all the rows of the mean with  $\gamma_j = 0$  are zero and all the rows and columns of  $\Omega$  with  $\gamma_j = 0$  are zero as well.

To retain the same form for the distribution of  $A$  as in the previous model, we make it so that, conditional on  $\gamma_j = 0$ , row  $j$  and column  $j$  of  $\Omega$  will all be set to zero (the resulting conditional distribution on the rest of the matrix is still Inverse Wishart with different degrees of freedom); and, conditional on  $\gamma_j = 0$ , row  $j$  of  $A_0$  will be all zeroes (the resulting conditional distribution on the rest of the matrix is still Gaussian). This gives rise to the following model:

---

<sup>3</sup>It is easy to see why we would refer to the  $\gamma_j = 1$  distribution as a “slab” when it is a uniform distribution, any picture of which in two dimensions will look like a flat slab of probability mass; other models like this are still called “spike and slab,” even where the full priors that are used for the  $\gamma_j = 1$  distribution not even particularly diffuse, just because they have the same structure otherwise.

$$\begin{aligned}
(86) \quad & \alpha \sim \text{Gam}(\alpha|a, b) \\
& \gamma_1 = 1 \\
& \gamma_j \sim \text{Bernoulli}(\tau) \text{ for } 2 \leq j \leq h \\
& A_{0,\gamma} \sim \mathcal{N}(A_{0,\gamma}|M_{0,\gamma}, S_{0,\gamma}, \mathbb{I}_h) \\
& A_{0,-\gamma} = 0 \\
& \Omega_\gamma \sim \mathcal{IW}(\Omega_\gamma|\Phi_\gamma, \lambda - h + \sum_{j=1}^h \gamma_j) \\
& \Omega_{-\gamma} = 0 \\
& G_0 = \mathcal{N}(A|A_0, \Sigma, \Omega) \times \mathcal{IW}(\Sigma|\Psi, \kappa) \\
& G \sim DP(\alpha G_0) \\
& A_i, \Sigma_i \sim G \\
& \underline{y}_i|A_i, \Sigma_i \sim \mathcal{N}(A_i^T \bar{x}_i, \Sigma_i)
\end{aligned}$$

Although it is certainly possible to construct a Gibbs sampler in which  $\gamma$  is updated conditional on the current values of  $\Omega$  and  $A_0$ , in practice such a sampler cannot move away from local optima for  $\gamma$ , because the principal effect of a change to  $\gamma$  is to propagate down to the likelihood (through  $\Omega$  and  $A_0$ ) and thus impact the assignment of observations to categories in the Chinese restaurant process in a way which is reflected somewhat subtly in  $\Omega$  and  $A_0$ . Thus the values of  $\gamma$  and  $z$  are sampled in block (throughout,  $h \times p$  location matrices subscripted with values of  $\gamma$  are reduced to contain only the rows with  $\gamma_j = 1$  and  $h \times p$  scale matrices subscripted with values of  $\gamma$  contain only the rows and columns with  $\gamma_j = 1$ ):

(87) Algorithm 2.

1. For  $j$  in  $2, \dots, h$ :

(a) For  $\gamma_j^* = 0, 1$ :

i.  $A^* \leftarrow A_{\gamma^*}, \Omega^* \leftarrow \Omega_{\gamma^*}, A_0^* \leftarrow A_{0,\gamma^*}, \Phi^* \leftarrow \Phi_{\gamma^*}, M_0^* \leftarrow M_{0,\gamma^*}$

ii.  $h^* \leftarrow \sum_{g \in \gamma_{-j}} g + \gamma_j^*, \lambda^* \leftarrow \lambda - h + h^*$

iii.  $\pi_{\gamma_j^*} \leftarrow \frac{|\Phi^*|^{\frac{\lambda^*}{2}}}{2^{\frac{h^* \lambda^*}{2}} (2\pi)^{h^*} \Gamma_{h^*}(\frac{\lambda^*}{2}) |\Omega^*|^{\frac{\lambda^*}{2}} \frac{h^*}{|S_0|^{\frac{h^*}{2}}} \text{etr} \left\{ -\frac{1}{2} \Phi^* \Omega^{*-1} \right\} \text{etr} \left\{ -\frac{1}{2} S_0^{-1} (A_0^* - M_0^*)^T (A_0^* - M_0^*) \right\}}$

iv. For  $i$  in  $1, \dots, N$  (where  $N$  is the total number of data points):

A.  $z_i^*$  is resampled as before by computing the posterior for each  $z^*|z_{-i}$ , substituting  $A^*, \Omega^*, A_0^*$  for  $A, \Omega, A_0$

B. New parameter values are sampled as necessary, conditional on  $\gamma_{-j}$  and  $\gamma_j^*$

C.  $\pi_{\gamma} \leftarrow \pi_{\gamma} \cdot \text{posterior}[z_i^*|z_{-i}]$  for the selected  $z_i^*$

(b) Sample  $\gamma_j$  from  $\langle \pi_0, \pi_1 \rangle$

(c) Set all  $z_i$  according to the  $z_i^*$  proposed under  $\gamma_j$

2. For each distinct  $z$ , let all  $\Sigma_z$  such that  $z_i = z$  be a sample  $\Sigma_z$  from an inverse Wishart distribution with parameters:

$$\begin{aligned} \Psi'_z &= \Psi + Y_z Y_z^T + A_{0,\gamma}^T \Omega_{\gamma}^{-1} A_{0,\gamma} - \left[ X_{z,\gamma} Y_z^T + \Omega_{\gamma}^{-1} A_{0,\gamma} \right]^T \left[ \Omega_{\gamma}^{-1} + X_{z,\gamma} X_{z,\gamma}^T \right]^{-1} \left[ X_{z,\gamma} Y_z^T + \Omega_{\gamma}^{-1} A_{0,\gamma} \right] \\ \kappa'_z &= \kappa + N_z \end{aligned}$$

3. For each distinct  $z$ , let all  $A_{i,\gamma}$  such that  $z_i = z$  be zero and  $A_{i,\gamma}$  be a sample  $A_{z,\gamma}$  from a normal distribution with parameters:

$$\begin{aligned} A'_{0,z} &= \left[ \Omega_{\gamma}^{-1} + X_{z,\gamma} X_{z,\gamma}^T \right]^{-1} \left[ X_{z,\gamma} Y_z^T + \Omega_{\gamma}^{-1} A_{0,\gamma} \right] \\ \Sigma'_z &= \text{the } \Sigma_z \text{ shared by the } z_i = z \\ \Omega'_z &= \left[ \Omega_{\gamma}^{-1} + X_{z,\gamma} X_{z,\gamma}^T \right]^{-1} \end{aligned}$$

4. Sample  $\Omega$  from an inverse Wishart distribution with parameters:

$$\Phi' = \Phi + \sum_{z \in Z} (A_z - A_0)^T \Sigma^{-1} (A_z - A_0), \quad \lambda' = \lambda + |Z|p$$

where  $Z$  is the set of all distinct  $z_i$

5. Set  $\text{vec}(A_{0,-\gamma})$  to 0 and sample the subset of  $\text{vec}(A_{0,\gamma})$  from a normal distribution with parameters:

$$\begin{aligned} \text{location} &= \left[ S_0 \otimes \mathbb{I}_{h\gamma} \right]^{-1} + \left[ \left( \sum_{z \in Z} \Sigma_z^{-1} \right)^{-1} \otimes \Omega_{\gamma} \right]^{-1} \left[ \left[ S_0 \otimes \mathbb{I}_{h\gamma} \right]^{-1} \text{vec}(M_{0,\gamma}) + \sum_{z \in Z} [\Sigma_z \otimes \Omega_{\gamma}]^{-1} \text{vec}(A_{k,\gamma}) \right] \\ \text{scale} &= \left[ S_0 \otimes \mathbb{I}_{h\gamma} \right]^{-1} + \left[ \left( \sum_{z \in Z} \Sigma_z^{-1} \right)^{-1} \otimes \Omega_{\gamma} \right]^{-1} \end{aligned}$$

6. Sample a value  $x$  from a beta distribution with parameters  $\alpha + 1, N$ . Then sample  $\alpha$  from a gamma distribution:

- with parameters  $a + |Z|, b - \log x$ , with probability proportional to  $(a + |Z| - 1)$
- with parameters  $a + |Z| - 1, b - \log x$ , with probability proportional to  $N(b - \log x)$

Note that, during the sampling step for  $z$ , none of the existing parameter values for  $A$  are resampled in  $A^*$  for the proposal  $\gamma_j^* = 1$ , even if the current values were sampled with  $\gamma_j^* = 0$ ; these values are still consistent with the proposal that  $\gamma_j^* = 1$ ; for  $\gamma_j^* = 0$ , where the current values of  $A$  have density 0 in general, only the offending rows are updated in  $A_0^*$  and  $A^*$ . For  $\Omega^*$  we are not so lucky: the reduced matrix would be ruled out under the full inverse Wishart prior. The resampling step for  $\Omega$  thus samples the full matrix (equivalent

to first sampling  $\Omega_\gamma$  and then sampling the rest from the prior, conditional on the sampled values of  $\Omega_\gamma$ , since with  $A_{0,-\gamma}$  and  $A_{-\gamma}$  set to 0 the relevant cells of the inverse Wishart scale matrix are unaffected by  $A_{-\gamma}$ , even though it is sampled from  $\Omega$ ); the additional values simply serve as a proposed value in order to evaluate the likelihood integral and sample new parameter values in the Chinese restaurant process. The idea, in general, is that we sample all the model parameters in block with  $\gamma$ , but bypass the full resampling steps for the numerical parameters until  $\gamma$  has been fully updated. The probability with which  $\gamma_j$  is sampled is almost (but not quite) the joint posterior probability of  $\gamma_j$ ,  $\Omega$ ,  $A_0$ , and  $z$ .

### 3.3.2 Experiment: Inuktitut revisited

To test this model we will use the same Inuktitut data as in Dillon, Dunbar & Idsardi 2013, with several basic questions in mind:

1. The model should reliably choose to set  $\gamma = 1$  for the uvular-environment predictor; does it? We will need to compare against a control predictor that should be set to  $\gamma = 0$ .
2. Does the model reliably set the effects of other environments to be active also, and, if so, how well does this align with manual estimates of the size of these effects in the data?
3. The inference in this model is  $2 \cdot (h - 1)$  times as complex as for the old model; we may have concerns about the performance of the sampler as we increase  $h$ .

4. The feature marking the presence or absence of a uvular environment might be thought to be more or less available to the learner than features marking the presence or absence of various different segmental categories in the environment. We should compare how good the learner is at detecting the effects of the individual segments (in this case [q] and []) instead.

### 3.3.2.1 List of sub-experiments

We now document the setup for each of the relevant experiments (numbered in sequence following the Dillon, Dunbar & Idsardi 2013 experiments to avoid confusion):

Experiment 4 We use the same uvularity predictor as before;<sup>4</sup> we add a second predictor drawn uniformly at random from {0, 1}. The data is as before and we manipulate the size of the data set again by upsampling.<sup>5</sup>

Experiments 5–7 As in Experiment 4, but we add the corresponding predictors for coronal (Experiment 5), velar (Experiment 6), and labial (Experiment 7) following consonants.

Experiments 8–10 We combine the uvularity predictor with two (Experiment 8), three

---

<sup>4</sup>We use a different uvularity predictor than before. The earlier experiments used a predictor which was 1 if there was either a uvular preceding or following. Here, I used other predictors as well, and, since their effects might not have been similar when preceding and following, I restricted consideration to their effects when following the vowel; the post-uvular environment was used rather than the two-sided uvular environment for consistency. Results for the original predictor were qualitatively the same.

<sup>5</sup>In these experiments, I upsampled simply by adding Gaussian noise, rather than applying kernel smoothing, which is slightly different in that the variance is not fixed (though it is still small). The new samples are in multiples of the size of the original data set (239). This is because we had previously attempted to preserve the rough proportions of each phoneme–predictor value combination in the data set (or in fact to manipulate it slightly). Here I wished to simply preserve the proportions to make the data as comparable as possible, but with multiple predictors, the size of the table of conditions increases, and the count in each cell decreases, thus making it more difficult to preserve the proportions accurately when increasing the size in non-integer multiples.

(Experiment 9), and four (Experiment 10) uniform random predictors, not correlated with each other.

Experiment 11 As in Experiment 4, but we use two predictors, one for  $[q]$  and another for  $[\bar{q}]$ .

### 3.3.2.2 Results

As before, ten chains were run, each with a different held-out 10%. Three sample sizes were used (see footnote 5). Ten thousand burnin samples were drawn, followed by a sample of one thousand at a lag of seven. The highest posterior sample was selected. A small amount of annealing was found to help avoid the proliferation of categories; the temperature was lowered from 1 to 0.1 on  $z$  by exponential decay through to the end of burnin. Table 3.3 summarizes the results.



Experiment 4 (uvular)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6	7	R	U			
239	0.70	0.61	0.84	0.6	0.4					0	1.0			
478	0.76	0.76	0.77	0.1	0.8		0.1			0	1.0			
717	0.73	0.88	0.62				0.9	0.1		0	1.0			
Experiment 5 (uvular and coronal)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6		R	U	C		
239	0.67	0.56	0.86	0.8	0.2					0.4	1.0	0.1		
478	0.73	0.74	0.74	0.2	0.5		0.2	0.1		0.2	1.0	0		
717	0.74	0.90	0.62				0.9	0.1		0	1.0	0.4		
Experiment 6 (uvular and velar)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6	7	R	U	V		
239	0.68	0.58	0.84	0.7	0.3					0.1	1.0	0.1		
478	0.76	0.75	0.77	0.1	0.8		0.1			0.1	1.0	0		
717	0.73	0.91	0.61				0.4	0.4	0.2	0	1.0	0.4		
Experiment 7 (uvular and labial)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6	7	R	U	L		
239	0.64	0.50	0.86	1.0						0.1	1.0	0.1		
478	0.73	0.73	0.75	0.2	0.6		0.1	0.1		0.1	1.0	0		
717	0.73	0.90	0.62				0.8	0.1	0.1	0.1	1.0	0		
Experiment 8 (uvular and two random)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6	7	R1	R2	U		
239	0.64	0.50	0.89	1.0						0	0.7	1.0		
478	0.75	0.68	0.85	0.4	0.6					0	0.1	1.0		
717	0.78	0.77	0.80	0.1	0.7	0.2				0	0.1	1.0		
Experiment 9 (uvular and three random)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6	7	R1	R2	R3	U	
239	0.65	0.51	0.89	0.9	0.1					0.7	1	0.9	1.0	
478	0.75	0.69	0.82	0.3	0.7					0.9	0.9	0.7	1.0	
717	0.78	0.76	0.81	0.1	0.9					0.7	0.9	0.2	1.0	
Experiment 10 (uvular and four random)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6	7	R1	R2	R3	R4	U
239	0.67	0.54	0.89	0.8	0.2					1.0	1.0	1.0	1.0	1.0
478	0.70	0.59	0.89	0.7	0.3					1.0	1.0	1.0	0.9	1.0
717	0.78	0.76	0.80	0.1	0.9					0.9	0.8	0.9	0.9	1.0
Experiment 11 (uvular stop, uvular fricative)														
$N$	$F$	Precision	Recall	$K = 2$	3	4	5	6		R	q			
239	0.77	0.78	0.77		1.0					0	1.0	1.0		
478	0.72	0.75	0.69		0.7	0.1	0.2			0.2	1.0	1.0		
717	0.65	0.80	0.55			0.1	0.9			0.1	1.0	1.0		

Table 3.3: Quantitative summary of Experiments 4–11, described in 3.3.2.2. See Table 3.1 for an explanation. The columns at the right tabulate what proportion of the chains set  $\gamma_j$  to 1 for each of the variables. The R, R1–4 variables are the random control predictors. U is for uvular; V is for velar; C is for coronal; L is for labial.

We return to our questions:

1. Does the model reliably choose to set  $\gamma = 1$  for the uvular-environment predictor?

Yes. Looking at both Experiment 4 and at all the other experiments including a uvular-environment predictor, (Experiments 5–10), the model consistently chooses to activate that predictor. In Experiments 4–7, with a random control predictor, this choice is clearly distinct from the choice of whether to activate the random predictor, which is activated rarely.

2. Does the model reliably set the effects of other environments to be active also?

(a) Coronal: No. In the runs with the most balanced  $F$ -scores (478 data points), which also have the highest proportion of runs with three categories, the coronal predictor is deactivated. In the runs under the larger data set, the coronal predictor is activated reasonably often; this may be related to the poorer alignment of these categories with the true phoneme categories.

(b) Velar: No. The pattern is the same as for coronal.

(c) Labial: No. As for coronal.

3. Does performance decline as we increase  $h$ ? Yes. The model correctly deactivates the random predictors consistently when there are no more than three predictors total (and even then only given that the category model is reasonably good: consider Experiments 5 and 8, where a random predictor is sometimes activated for two-category solutions). This suggests a very small limit on the number of predictor variables that can be meaningfully selected among.

	Velar		Coronal		Labial		Uvular	
	$F_1$	$F_2$	$F_1$	$F_2$	$F_1$	$F_2$	$F_1$	$F_2$
a	-38	12	-20	67	23	24	47	-69
i	-19	128	-37	62	16	-101	101	-376
u	-23	31	-39	75	-52	42	57	-102

Table 3.4: Difference between the Inuktitut vowel mean conditional on a particular following consonant place and the mean elsewhere, for the four different places of articulation.

4. Can the model detect the effects of the individual uvular segments ([q] and [ʁ]) instead of their disjunction? Yes. Compare the results of Experiment 11 to Experiments 5–8, where, out of three predictors, only one was reliably activated; here, both the [q] and [ʁ] predictors are activated.

To investigate these results somewhat further, we consider the effects of coronal, velar, and labial environments. Should we expect to find an effect of either of these two types of segments? If so, how do the model’s estimates compare to reality? The latter question is moot. For the coronal model, further investigation revealed that none of the three-category models had the coronal predictor activated; there was only one three-category model with the random control predictor active. The same held for the velar model, except that the random control predictor was active in two three-category models. For the labial model, none of the three-category models had either the labial or the random control predictor active. Thus no comparison could be made. The data itself shows that the means conditional on these environments are substantially more similar to the means elsewhere than in the case of uvular following consonants. See Table 3.4.

The second issue is the model’s performance as we increase  $h$ , which is poor (with respect to the selection of predictors) as we go past three total. A reasonable explanation

is that the sampler might not move efficiently through the high-posterior regions of the hypothesis space, and in particular might either get completely stuck in a locally high-posterior region, or just move very slowly. This is definitely the problem: analysis of burnin shows the three-random-control model leaving the state with all  $\gamma_j$  set to one on an average of 0.1 percent of samples, scattered throughout burnin, despite the higher posterior value of these samples. At least for the Inuktitut data, this does not hinder the exploration of  $z$ : the high-posterior mixtures are generally reasonably good approximations to the three Inuktitut phonemes; the use of contextual predictors can improve the alignment to the true categories, but never seems to substantially worsen it. More efficient methods for sampling  $\gamma$  should be investigated; we may be able to take this fact into consideration at least in some cases.

### 3.3.2.3 Discussion

Given the above discussion of the linear phonetic transform hypothesis, we said nothing to imply that every contextual dimension would necessarily have an associated phonetic transform. This is what we are forced into, given the model presented in the previous section. In this section, we have introduced a model that learns which contexts trigger phonetic transforms. It can be said to be “learning the environments” for these shifts. Although this characterization might be counter to the intuitions of some, because we provided the model with a list of potential triggering environments, it should be noted that the existence of such a list (finite or infinite) is implicit in any model that says there is a discrete set of transforms, and any debate about whether the list “exists” can only be

a debate about how that information is represented (though representational questions are obviously highly relevant to the evaluation measure).

The role of the Bayesian Occam's Razor in a model like this is to guard against selections of less sparse solutions for the sake of only small improvements to fit (e.g., to capture the negligible "effects" of the randomly generated predictors), in this case via the dimension of  $A_0$  and  $\Omega$ . This yields a testable prediction, and the most careful empirical test of this part of the theory would look like this:

- (88) Find a satisfactory experimental methodology for training people (presumably infants) on novel phonetic categories, and assessing the resulting phonetic maps after training
- (89) Construct a data set which exposes subjects to sequences with phonetic tokens appearing in different environments, ensuring as well as possible that the environments provide salient cues which are not themselves subject to category learning (perhaps by tightly controlling their variance)
- (90) Manipulate, across conditions, the size of the phonetic effects of two different predictors; a critical comparison is between a condition with very different effect sizes for two predictors, one predictor's effect being very small, (different condition), and one with two comparable, reasonably large effect sizes in different directions (same condition)
- (91) Evaluate the integrated likelihood for the set of phonetic category models with and without each of the predictors

(92) The subjects' resulting phonetic maps should reflect net zero influence of the small-effect predictor in the different condition, while the phonetic maps in the same condition should reflect influence of the predictors in proportion to, and in the direction of, their acoustic effects

This experiment is reasonable but is beyond what psycholinguistics can currently achieve. First, there is one standard procedure for training on new phonetic categories (Maye, Werker & Gerken 2002) but it is unreliable (Peperkamp 2003, Gulian, Escudero & Boersma 2007). Second, although there is a standard method for probing listeners' phonetic maps (compare discrimination abilities for pairs of sounds in different parts of the phonetic space), fine-grained studies of what the changes in this measure due to contextual effects look like have not been done, to my knowledge. There are also unexplained phenomena found in this measure, such as effects of order of presentation of stimuli (Kuhl 1991). Third, discrimination measures in this experiment would be discrimination in context, which allows listeners to discriminate on the basis of the context, interfering with the probe of the phonetic map for the vowels; thus a different way of probing the phonetic map altogether would be preferable. Finally, although we could of course base likelihood computations off the Gaussian models used here, some predictions based on empirically grounded measures of "goodness" for a given phonetic mixture and set of data points would be better. Although the model's predictions are clear, therefore, they are difficult to test at present. Nevertheless, until this model is embedded in a larger model (including phone segmentation, word segmentation, categorical phonological grammar, and so on), it serves as the best demonstration possible that the phonetic transform hypothesis is a

feasible one from the point of view of learning.

### 3.3.3 Experiment: sex and gender differences

The idea behind the model presented in this chapter is more general than allophony, in two different ways, which I will briefly demonstrate now.

As outlined at the start of the chapter, we assume that context representations drive the effects of phonetic transforms. However, the mechanics of the mixture of linear models do not rule out these scalars being real numbers. This says nothing of the cognitive architecture, of course, but it means that it is also possible for us to construct models where the predictors are continuous, not discrete values (presence or absence of a certain environment)—if we like to think about the (epiphenomenal) surface phones, then in that sense this makes the inventory of phones nonenumerable.

Second, the predictors do not need to be anything to do with the segments in the immediate context; they could be any other property of the observation. One problem which has exactly the same structure as the allophone problem is the problem of indexical differences in sociophonetics: characteristics of the speaker that are realized in the phonetics, including both the aspects which are physical, such as sex differences due to vocal tract length, and those that are the result of applying sociolinguistic knowledge in production, such as (strict) gender differences (Foulkes, Scobbie & Watt 2010).

In this section, I demonstrate the applicability of the model to these more general cases by modelling a small vowel system with speaker variability, including both categorical and continuous predictors.

Experiments 12–13 English corner vowels, /i/, /e/, /u/, taken from Hillenbrand, Getty, Clark & Wheeler 1995. 347 observations consisting of the first three formants, taken at a steady state, with three predictors:  $f_0$  (fundamental frequency), sex (male or female), and age (adult or child). See Figure 3.5. Estimation of a regular mixture of Gaussians (Experiment 12) and the current model (Experiment 13).

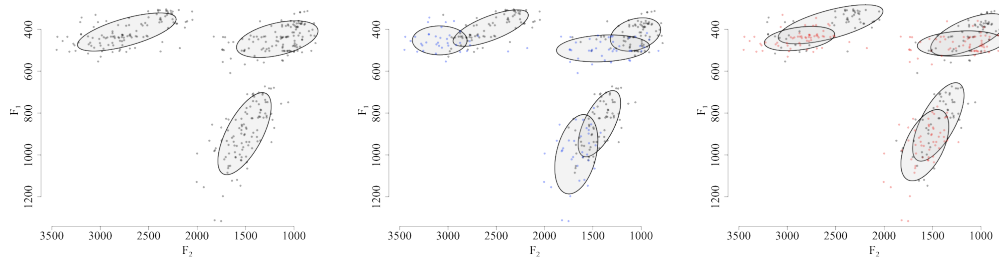


Figure 3.5: English corner vowels. Shown here with no division by predictor; with division by age (children blue); with division by sex (females red); from left to right.

As before, 10000 burnin samples were drawn, before taking a sample of size 1000 at a lag of 7, with ten chains each, with non-overlapping 10% subsets held out. The highest posterior sample was taken to be representative of a chain. For this model, the qualitative predictors were set to  $-1$  and  $1$ . Table 3.5 shows a quantitative assessment of the results.

Experiment 12 (English, mixture of Gaussians)										
$N$	$F$	Precision	Recall	$K = 2$	3	4				
347	0.95	0.98	0.93		0.7	0.3				
Experiment 13 (English, variable selection)										
$N$	$F$	Precision	Recall	$K = 2$	3	4	R	Sex	Age	$f_0$
347	0.99	0.99	0.99		1.0		0	1.0	1.0	1.0

Table 3.5: Quantitative summary of results of mixture of Gaussians versus variable selection model on English corner vowel data.

The variable selection model reliably correctly rejects the control predictor, indi-



cating that there is no technical issue with spurious solutions as there was before. All the other predictors are consistently selected; this is consistent with the fact that the predictors all have substantial effect. An example model is shown in Figure 3.6.

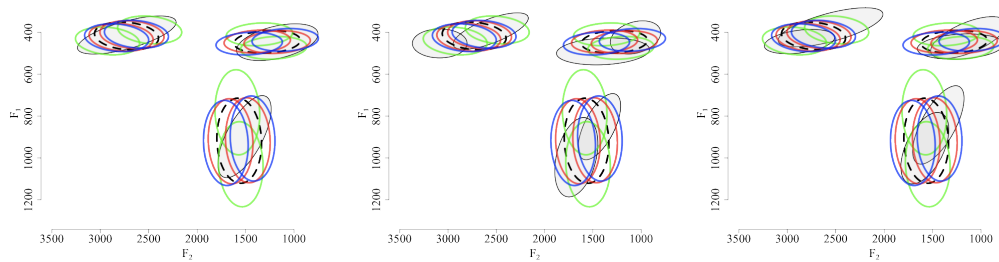


Figure 3.6: English corner vowels, one fitted model. Filled ellipses are Gaussians fit in a supervised way to individual sub-categories: no division by predictor; division by age; division by sex. Thicker lines are realized (epiphenomenal) categories: the dotted line is the prediction when the categorical predictors are coded as 0 and  $f_0$  is at its mean value for that category; green ellipses show the effect of  $f_0$ , with predictions at the 25th and 75th percentile  $f_0$  for the given category; red ellipses show the two predictions for sex; blue ellipses show the two predictions for age.

The idea that sociolinguistic and other talker differences could be handled using exactly the same cognitive apparatus as allophony is new; here I do not explore it in great depth, but I believe it is viable, and, given a phonetic transform view of allophony, there is nothing against uniting these two things. For the skeptical modularist worried about extralinguistic information contaminating the grammar, is important to remember that (i) the information triggering the transform is extralinguistic, and is provided to the linguistic system in order to decide how to process the input (in a broad sense “which grammar” to use, although that characterization applies equally to allophonic conditioning)—the phonetic representation coding the transform itself is not extralinguistic; and (ii) normalization for sociophonetic factors is just a fact (Evans & Iverson 2004)—so it needs to happen somewhere along the line in the phonetic system. This model does that. Since the structure of

the contextual and the indexical variability problems is so similar, the idea that the phonetic computations fundamentally change their character to deal with one as opposed to the other seems like it should not be the null hypothesis.

Beyond this suggestion, the results indicate that learning relies to some extent on talker normalization: there is improvement in the resulting solutions when the talker-level predictors are added to the model, as compared to when they are omitted. Of course, there are other sources of information besides the simple correlation of some aspect of the data with a talker variable which might cue the learner to find ways to ignore variability in phonetic realizations: hearing words which are identifiably the same pronounced quite differently by different speakers would presumably provide some information about speaker-level variability. However, using lexical information like this does not solve the problem of talker variability by itself. In principle, hearing a few high-frequency words pronounced systematically differently from what is expected under the listener's existing phonetic model could be a strong cue to the existence of a transform at the talker level; but learners do need to be able to generalize from this word to a model that will help them make sense of the stream in the future, and so it would not be enough to simply say that learners can fill in the category identity from context. Rather than just treating talker variability as a kind of mispronunciation, this model allows for sociophonetic and other types of talker variability to be treated as systematic.

### 3.4 Proposed model: learning with features

#### 3.4.1 Background: features, geometries, and the contrastive hierarchy

In this section, a model is proposed (but not implemented) which learns phonetic categories that are cognitively structured using binary valued distinctive features. I review the reasons why one might want to make such a move now.

Recall that, beyond phonetics, the ultimate problem of phonology is how the brain links phonetic representations with memory (the lexicon); the elements of the lexicon have some information relevant to phonetics (in addition to the syntactic and semantic information that links a word with its grammatical properties and its meaning); the information in a single lexical item is usually thought to consist of a sequence of “segments”; the information in a single segment is usually thought to consist of feature–value pairs, and the values are usually thought to be binary. The idea behind the use of binary valued features in phonology is thus that they are the smallest units of lexical storage.

A given set of feature value pairs induces an equivalence class, namely, the set of all objects with that set of feature–value pairs. For example, the vowel /i/ is often described as being [+high][−back]. All phonetic realizations of /i/ belong to this class. The vowel /u/ is then usually described as being [+high][+back]. All phonetic realizations of /u/ belong to this class. Notice how this is different from the models we have been using up to now. In particular, there is a logically possible class [+high] to which all phonetic realizations of /i/ and /u/ belong; the mixture models we have been using treat each phonetic category as atomic, but feature-based models treat a phonetic category as complex; valued

features are the atoms.

There are two crucial things here: first, segments belong to equivalence classes like /i/ and /u/ (and the assumption is often that nothing beyond these equivalence classes can be coded in memory, although this appears to be wrong: see, for example, McMurray, Tanenhaus & Aslin 2002). Second, segments are cross-classified: they belong to multiple equivalence classes. Even before grammar was given an explicitly mentalistic interpretation, and these equivalence classes were required to be psychologically active, many grammarians developed descriptions of phonological patterns that also crucially had to cross-classify segments. For example, Pāṇini's 6th-century BCE grammar of classical Sanskrit contains the "Shiva sutras," which is a table of classes of Sanskrit phonemes set out as verses: each verse is essentially a list of phonemes and a dummy phoneme at the end marking a class that all those phonemes belong to. Rules in Pāṇini's grammar then make reference to sequences which are of the form  $pP$ , where  $p$  is some phoneme symbol, and  $P$  is one of the dummy phonemes; this means the set of all phonemes starting from where  $p$  appears in the table, up to the end of the verse that  $P$  ends, where the "table" is arranged by reading the verses left to right, from the first verse to the last. For example, all the sonorants come in earlier verses than the obstruents, so to make reference to all the obstruents, one simply writes the symbol for the first phoneme that appears in the first verse with obstruents (which happens to be [d<sup>h</sup>]) followed by the symbol for the dummy phoneme at the end of the last line of obstruents (which happens to be [l]); however, one can make reference to narrower subclasses of obstruents, like the voiced unaspirated stops, by starting with the first such obstruent, [d], (from the second verse of obstruents), and ending with the dummy phoneme ending that verse, [], because that is the only verse containing

voiced unaspirated stops; see Kiparsky 1991. There are even more classes, too, not just the classes corresponding to sets of verses, because one does need to start at the beginning of the verse—but the point is merely that (i) there are finitely many equivalence classes of segments—Pāṇini does not make reference to gradient phonetic detail—and (ii) the equivalence classes overlap.

The reasons for (i) and (ii) in describing phonological patterns are clear: there are obvious patterns in pronunciation which seem to be of the form “whenever a segment of equivalence class X is in environment Y, it is actually pronounced as equivalence class X’”—the idea of a change from one form to another being backed up traditionally by morphological evidence about what the “basic” segmental form of a morpheme is—or of the form “segments of class X can never/only appear in the environment Y”; Pāṇini’s grammar, for example, describes Sanskrit with the rule “obstruents [‘[d<sup>h</sup>][l]’] change to voiced unaspirated stops [‘[d][ ]’] before voiced stops [‘[d<sup>h</sup>][ ]’],” making reference only to equivalence classes, and furthermore to equivalence classes cross-classified using features. As far as any grammarian knew, no reference to phonetic detail was necessary in describing these patterns, ever: the pronunciations after one segment is changed to another are just like the corresponding pronunciations when they are not the result of a change (“underlying” in modern cognitive terminology), or when they are changed from some other class of segment. Crucially, for many processes, this turns out to be true (though not all, of course, as discussed above: see Chapter 5 for more discussion). The existence of such processes is the definitive linguistic argument for these equivalence classes, and, to the extent that they must be cross-classifying to describe all such processes in a given language, for distinctive features.

Features can be leveraged for phonological patterns so that they have another benefit beyond defining overlapping classes, (iii), correspondence—so, in Sanskrit, the obstruent [d<sup>h</sup>] changes to the corresponding voiced unaspirated stop, [d], not some other one. In modern linguistic theory, this process would be described as replacing the feature value for [voice] on those segments with [+voice], replacing the [continuant] value with [–continuant] (for “stop consonant”), and replacing the feature value for [spread glottis] with [+spread glottis]. All the other feature values remain the same, and so different segments can be made to change to different corresponding segments, by making reference to changes to particular features. There do indeed seem to be many processes that work this way—putting segments in correspondence according to independently motivated featural classifications—and, if one needed an argument for allowing correspondence, that would be it (but one does not, since correspondence does not require any extra mechanisms under modern feature theory).<sup>6</sup>

Moving forward to the twentieth century, the Prague school phonologists—most notably, Jakobson and Trubetskoy—like all other phonologists, had devices for (i), (ii), and (iii), and these are the ancestors of modern feature theory. There are several notable things about this literature. One thing which was not really new was that phonological features were required to have some phonetic meaning: [+high] needed to mean something having to do with the position of the tongue when pronouncing a vowel, different, at least in a relative sense, from [–high]—or to do with some acoustic or auditory property (in which case the feature would have some other name), or both at once. In fact, it had been true

---

<sup>6</sup>In Pāṇini’s grammar, this fact, correspondence, is actually not handled by exactly the same device used to define the two classes, obstruent and voiced obstruent—instead, by the “closest place of articulation” clause, 1.1.50, which does not use the Shiva sutra verse features—but this is, at any rate, another cross-classification.

even for Pāṇini that phonological features were in some way phonetically grounded, as the Shiva sutras were clearly organized according to well-understood phonetic classifications of Sanskrit phonemes, but, in contrast with the Prague school (and even within the Prague school sometimes), some early twentieth century phonologists seemed to endorse the idea that the features were motivated entirely by the phonological patterns they capture—so that, say, the existence of the class of obstruents in Sanskrit would be motivated only by the fact that they are a set of segments that all undergo a change to voiced stops; this point of view will be discussed later, but suffice it to say that, at least for Jakobson, features were clearly phonetic, and were required to play a role in both speech perception and speech production.

The strong idea is carried forward into mainstream phonological theory today: there is a fixed set of binary features, a set of substantive phonological universals. There is some difficulty posed by this, of course, since the pronunciation of, say /i/, will always be slightly different across languages, and so both in production and in perception it seems quite reasonable to say that, rather than each phonological feature having a fixed phonetic interpretation, there is a fixed set of biases towards forming various different phonetic kinds of features. A strong version of this idea is implicit in Jakobson, Fant & Halle 1952; Jakobson & Halle 1956: these works attempt to delimit all the possible phonological features and associate each with a description of their acoustic, articulatory, and auditory characteristics. The binary feature [strident], for example, has at its positive value spectra with a “random distribution of black areas,” due to “turbulence at the point of articulation,” and identification of manipulated speech sounds as being the [+strident] element of a pair of stops in perception experiments (which according to Jakobson, Fant and Halle

is the affricate, so, [t] as opposed to [tʃ]) is most reliable when the duration of the sound is longer. At its negative value, (which Jakobson, Fant and Halle call mellow), it gives rise to “spectrograms in which the black areas may form horizontal or vertical striations,” and in production lacks the “supplementary barrier that offers additional resistance to the air stream” that the corresponding [+strident] will have, such as obstruction from the lower teeth or uvula. This gives rise to a strong suggestion that there is a one-to-one mapping between auditory feature detectors tuned to certain acoustic properties, on the one hand, and gestures, on the other, which some theories of speech perception pursue (Fowler 1986); others propose that there is a specialized system early in perception that converts percepts into motor representations (Lieberman & Mattingly 1985), so that all phonological computations can be done over production features, while still others propose that perceptual and production phonetics are linked only by a prediction from a “forward model” to predict what the percept should be for a given lexical item, stored in terms of some production-oriented features (Stevens 2002). Finally, some authors have argued for a mix of perceptual and production features in the lexicon (Ladefoged 2005). Note that this phonetic content in binary phonological features is asserted in spite of the fact that at some level the phonetics definitely contains graded information. Our assumption is that, even if the phonetic systems do have binary features in them, learned, language-specific computations can be done over the graded information, not only over binary features; see Chapter 5 for discussion of the empirical issues.

In addition to making phonological features play a key role in perception and speech production, Jakobson 1941 places even greater empirical demand on features. He claims that the oppositions they set up delimit how infants learn: infants go through a sequence



of increasing complexity in what sounds they can pronounce, first learning to make the distinction associated with one feature (say, consonants versus vowels), then refining this by adding another feature (according to Jakobson, high versus low on the vowel side, and oral versus nasal on the consonant side). There is a fixed set of binary phonetic features, and an ordering is stated over these features; children's language acquisition follows this ordering, and makes a progression following these distinctions and not other ones. The reverse order was supposedly the order that distinctions were lost in aphasia. There is little empirical support for Jakobson's claims (see "The Acquisition of Phonological Inventories" for a review) but, more generally, it has been common throughout the literature to make phonological features act as a set of fundamental units common to various systems and processes and have broad effects in all of them.

Given that the understanding is that (at least for a given language), there is a finite set of phonological features available, underspecification is the idea that a segment can be marked with some, but not all, of these possible features. There are two different ideas that sometimes get called "underspecification." One is that features do not need to be valued at all, and, instead, they are marked as present or absent; the other is that features do need to be valued, but not all the feature–value pairs need to appear. The first theory, a theory that features are privative, is fairly consistently understood to imply that the phonological grammar cannot change or condition on unspecified features, because they are not present in the representation to begin with. However, it is often unclear what the difference between this theory and the binary value theory is as far as the phonetic interpretation of features goes: what is the difference between saying that the position of the tongue, or the value of the first formant, in a vowel, can be specified as [high], or contain no such informa-

tion, as versus the claim that the phonetic specification can either be [+high] or [−high]? In either case, there must be one specification that corresponds, phonetically, to high vowels, and another that corresponds, phonetically, to low vowels. One answer is suggested by Lahiri & Reetz 2002: features may be underspecified in lexical representations, but those same features may be perceived—that is, the perceptual system may make use of them and attempt to look up words in memory using those features: the feature [coronal], which characterizes sounds like [t] and [n], is often thought to be underspecified in all languages. The prediction is that, in word recognition, other features, like [labial], which characterizes sounds like [p] and [m], should be able used as successful search probes for words with coronal consonants: the German word *Düne*, “dune,” should be accessed when a listener hears the non-word \**Düme*, but the word *Schramme*, “scratch,” should not be accessed when a listener hears the non-word \**Schrane*. For some empirical evidence like this, see Cornell, Lahiri & Eulitz 2011; Lahiri & Reetz 2010; Scharinger, Lahiri & Eulitz 2010; Scharinger & Lahiri 2010.

The second idea of underspecification, in contrast, does not just say that presence/absence is just the way that positive and negative feature values are represented (with whatever consequences that has); rather, the representation must contain a value, positive or negative, when it contains a given feature, but it does not need to contain every feature. There are thus three values, in the sense of three possible states, for a given feature, which sometimes get called +, −, and 0; again, though, 0 values, when they mean “unspecified,” generally imply that the feature is invisible to phonological operations. As for phonetics, the same perceptual predictions could be made, but, in this theory, because underspecification implies that the feature is never specified at all, positively or negatively, there is

another possibility, suggested by Cohn 1993; Dyck 1995; Keating 1988: lack of specification implies greater variability in production, so that, for example, since Russian [x] is underspecified for the feature [back], but [], according to Keating, is specified as [+back], [x] has greater front–back variance in its pronunciation, in terms of the position of the tongue. Both theories can be combined, too: some features may be privative, others binary.

The notion of contrast is also relevant to much of phonological feature theory. Dresher 2009b argues that the Contrastivist hypothesis (Hall 2007) is to be found, implicitly or explicitly, throughout the phonology literature: the idea that only a subset of the features that are needed to fill in all the relevant phonetic information will be able to be altered or conditioned on by phonological grammars, namely, those features which are contrastive. The notion “contrastive” is usually understood with respect to the inventory of segments (specified as sets of feature–value pairs), and in particular the inventory of segment classes that are used somewhere in the lexicon. These set up the possible contrasts: for example, Inuktitut has a three-way contrast between /i/ ([+high][−back]), /u/ ([+high][+back]), and /a/ ([−high][? back]). The reason it is difficult to say whether /a/ is specified as + or − for the [back] feature is not only because its phonetic realization is actually fairly central, (and so therefore neither clearly back nor clearly front), but also because it is the only low vowel that ever needs to be coded lexically, and, therefore, whichever of the two categories [−high][−back] or [−high][+back] the lexicon actually stores, it does not seem to ever use the other. We say that the feature [back] is not contrastive for low vowels in Inuktitut, and the Contrastivist hypothesis would therefore predict that the feature [back], whatever its value might be, is invisible to the phonological mapping when it appears on

a segment in combination with [−high].

Unfortunately, there is some difficulty in jumping from an idea of contrasts between segments to contrastiveness of features in this way, as Dresher points out: without the question about the phonetic backness of /a/, we could just as easily have said that there is a two-way height contrast between /u/ ([+high][+back]) and /a/ ([−high][+back]), but that, since there is no low front vowel corresponding to /i/, only the feature [back], but not the feature [high], is contrastive for /i/. Dresher appeals to an ordering of features to fully determine contrastiveness of features: one asks whether a particular feature is contrastive for given all the contrastive specifications of the features previous in the order, so that [back] is not contrastive for /a/ if  $\text{high} < \text{back}$ , while [high] is not contrastive for /i/ if  $\text{back} < \text{high}$ . The idea is the same, however: by some criterion having to do with what segments are in the inventory—a criterion which will be dependent on setting a feature ordering, according to Dresher—the specification of a given feature can be determined to be contrastive, as opposed to redundant, for some subset of the inventory. The Contrastivist hypothesis is often reduced to some version of contrastive specification (Steriade 1987, Dresher 2009b): non-contrastive features are always underspecified, and this is why phonological grammars cannot see non-contrastive features—they are absent.

Throughout this section, we have seen that different empirical demands have been placed on distinctive features: are they there to explain phonological patterning only—which, in the cognitive view, makes them lexical units that the phonological grammar manipulates—or are they also part of the system of phonetics, either in perception or production or both? This question is frequently reduced to the following: are phonological features merely classificatory, grouping segments together for the purpose of stating the

patterns captured in phonological grammars, or are they (also) phonetic, specifying particular information relating to perception or articulation?<sup>7</sup> Chomsky 1964 attributes the first view to Bloomfield, and argues instead for lexical information to be grounded in universal phonetics, building on Jakobson's ideas.

The question is frequently further reduced to the following: are phonological features learned, or “emergent,” (with no bias towards one type of feature or another coming from UG) or are they “innate” (with substantial bias towards certain features, like vowel height, consonant place of articulation, and so on, with a bit of fine tuning learned for each language)? The reason that the “weak versus strong” learning bias question is often strongly linked to, or even equated with, the classificatory versus phonetic question (for example, Morén 2007, Mielke 2008) is that, first, sound patterns differ widely from language to language, and so the classification of sounds for patterning purposes must be different across languages, thus learned, thus totally arbitrary (this last step in the reasoning is not justified by itself, but this is the argument that Mielke makes in favor of emergent features). The second part of the link seems to be that, since phonetic systems are limited by the auditory and articulatory systems, any phonetic features would need to be fairly tightly constrained in their content; this is true up to the “any”—one can also

---

<sup>7</sup>The understanding is, furthermore, almost always that features group segmental categories, like [i] and [u]. One line of argument against the classificatory view, therefore, would be that the basic facts about the segmental categories used in language cannot be learned, or even properly described, without reference to phonological features. The arguments above about how features must have some cognitive use in the phonetic system, if correct, strongly suggest an argument like this: if perceptual cognition works using features, and segments per se are not used in perception to at all, then in order for the classificatory view to be right, lexical storage would need to first wipe out all the featural information and turn the output of the phonetics into atomic categories, only to recreate a (different) system of features, and this would seem unlikely given how often phonological feature systems recapitulate phonetically grounded classes. If a phonetic category learning model doing featural analysis of the phonetics to form categories, as versus one treating categories as atomic, could be shown to be better in some way, it would be another suggestive argument of this kind.

imagine features that are phonetically grounded but learned in order to better encode the phonetic contrasts in a language. (Morén’s argument is that, since sign language and spoken language use different articulators, features must be learned: apart from relying on this assumption that features could not be phonetic and also be quite plastic, this also rests on the assumption that there could not simply be two different types of features used in the phonologies of sign versus spoken language.)

To sum up: distinctive features are dimensions of cross-classification for speech sounds. Today, they have a standard interpretation as basic units of lexical storage within segments. They are usually understood to be either binary (either “present/absent,” i.e., “specified/underspecified,” or else “+ value/– value”) or ternary (“absent/present with value +/present with value –”). Most phonological theories also use features to set up correspondences between segments. There is support for the equivalence classes and correspondences they set up from phonological patterning (again, see Chapter 5 for more discussion of the evidence for true equivalence classes in phonological processes). They are also understood to play a role in speech perception and in speech production: for perception, some authors take them to be fundamental units, while others take the segment or syllable to be fundamental, but still take features to be lexical units which can influence perception. For production, researchers agree that production systems need to make reference to the various different motor dimensions of a segment, and that phonemes *per se* are not fundamental cognitive units in production; theories of perception also recognize the fact that the auditory system is multidimensional, and differ over how the perceptual dimensions are connected with production dimensions, and whether the systems share a common set of binary features (that is, whether the two sets of are linked by a very strong

homomorphism preserving the precise coding of all the lexical items they are associated with). Finally, different views of the role of features in phonetics correspond to an active debate in phonological theory, that of classificatory versus phonetic features. Minimally, however, there is good reason to think that that some cross-classification into equivalence classes is necessary to handle phonological patterning, and, since binary or binary+underspecification type features are the standard way to do this, it is worth asking how we would incorporate this into a phonetic category learning model, and, thus ultimately, into a phonetic grammar learning model.

### 3.4.2 Background: Bayesian category models with features

Many statistical models also use binary valued features in a similar way to phonological feature systems. This is different from saying that there are many statistical models for binary- or categorical-valued data—it is generally implicit in any phonological feature theory that, if any of the phonetic values for features need to be learned, then the relevant “data” (input in the auditory system) is not categorical. Instead, what it says is that these models learn binary-valued parameter vectors from potentially continuous-valued input. In particular, there are a number of category models—which is to say, mixture models—in which each category is in some sense “made up of” a collection of categorical feature–value pairs.

To see this, start with the Chinese restaurant process model—which does not have this property—considering Algorithm 1: the indices (category labels)  $z_i$  are categorical in the sense that they are drawn from from an enumerable set. There is no reason that the

index set they are drawn from had to be the integers; the algorithm would have worked exactly the same if the indices had been Unicode symbols: Q, ♖, ♣, and so on. In this model, a category is uniquely associated with a single element of an index set  $\mathcal{U}$ , an element which is in turn associated with a parameter vector  $\theta$ , from a different set  $\Theta$  (in our case  $\langle A, \Sigma \rangle$  pairs). Specifying a particular value of  $\theta$  fills in the specification of a likelihood function connecting the model with the input (it gives us a “component of the perceptual map” in our case). The association between elements of  $\mathcal{U}$  and  $\Theta$  is arbitrary.

Consider now how this differs from the phonological feature systems described in the previous section. A single segmental storage unit (category) is associated with a set of feature–value pairs. If all the features are binary (or  $n$ -ary, etc) then we can say that the feature *values* are all drawn from a single set  $\mathcal{U}$ , but each category is represented, now, not by a single value  $u \in \mathcal{U}$ , but by a mapping  $c : \mathcal{F} \rightarrow \mathcal{U}$ , taking elements of some set of “possible features” as input (we discussed features like [high] and [back] above: these would be two different elements of  $\mathcal{F}$ ). Put aside for the moment the fact that the features (the elements of  $\mathcal{F}$ ) are understood to have content (that is, there “is” a feature [high] that is different from the feature [back] not because of the values it can take, but because of something about what it “means” to be [+high] as opposed to [+back]). There is a still more fundamental difference between the CRP model, known as a *latent class model*, and the feature model of lexical storage: categories are complex, not simplex. A single observation (if we are thinking about observations of individual segments) must be paired with more than just a single element in order to be considered a “member” of a particular category. In the statistics literature, models in which a single “category” is actually a complex object, are called *latent feature models*. We will review some of



these models now. However, before we continue: it is important to pay attention to the fact that these models, unlike the feature model of lexical storage, almost always work under the assumption that features do not have a priori content; a particular category will be represented in these models as just a *sequence* of feature values, not a mapping from features to values. A sequence can of course be *seen* as a mapping  $c$  in which the domain  $\mathcal{F}$  is the integers, but, for the purposes of these models, it can just as easily also be seen as a mapping in which  $\mathcal{F}$  is the Unicode table—the order that features are in in a model like this is just a way of keeping straight distinct items, not a way of giving them content; there is no sense in which models like this could “have” a feature [high] or [back] or anything along those lines, because the features are interchangeable. We will see this spelled out right now, and we will return to it later on, and discuss how to extend these models to obtain more constrained latent feature models.

To begin with an example: the standard Bayesian latent feature model is the *Indian buffet process (IBP)* due to Griffiths & Ghahramani 2006. In this model, the number of features is infinite and values are binary. The sampling scheme set out by Griffiths and Ghahramani is very similar to Algorithm 1 above, except that each observation must now be associated with, not a single category, but a vector of features; rather than evaluating, for each existing simplex category, the probability of assigning the observation to that category, plus some probability for adding a new category, we evaluate the probability of assigning each feature the value 1 for this observation, and, with some probability, set some number of new features to 1. In the IBP, the probability of setting feature  $k$  to 1 for a new observation (conditional on feature specifications for the previous observations) is  $\frac{N_k}{N}$ , where  $N_k$  is the number of previous observations with  $k = 1$ , and  $N$  is the total number of

observations (including the new one), weighted by the likelihood under that specification, as in the DP; the distribution on the number of new features used is Poisson with rate parameter  $\frac{\alpha}{N}$ , where  $\alpha$  is a hyperparameter, while the unconditional distribution on the total number of features used is Poisson with rate parameter  $\alpha$ , meaning that we should think of  $\alpha$  as the average number of features used by any given observation. Similar to the Chinese restaurant process, the “add to existing” probabilities increase in the number of other observations with the property, only, here, the property is not “belongs to the given category” but rather “has the given feature as 1.” Unlike in the Chinese restaurant process, however, the probability of creating a new (previously unobserved) category is not limited to the probability of adding new features: it is possible to create a new category out of old features, simply by selecting a combination that has not been added before.

Latent feature models thus decompose categories. A “category” in a latent feature model like the IBP is some combination of binary feature values; but remember that categories need to be associated with parameter values to do anything. If categories are decomposed in a latent feature model, it would seem to suggest that we must have decomposed the parameter values as well—after all, if the new feature combination  $f_1 = +/f_2 = -$  was not required to have anything in common, via the feature  $f_1 = +$ , with the category  $f_1 = +/f_2 = +$ , then the practical value of decomposing the categories using features becomes much less obvious. What is the equivalent of sampling a new parameter value in the IBP? In general it depends on what the features are modelling (they do not need to be modelling phonetic categories), but the key is that there is some parameter value  $\theta_f$  associated with each feature  $f$ , and there is some structure in the set of likelihood functions which is preserved under conjunction of features, so that  $\lambda[f_1 \wedge f_2 \wedge \dots] = \lambda[f_1] \odot \lambda[f_2] \odot \dots$ . The

example that Griffiths and Ghahramani use is of multivariate Gaussian likelihoods with locations differing as a function of the category—rather like our own phonetic category models—and with  $\odot$  being the operation that simply sums the locations of the Gaussian likelihood functions, so that, for example,  $\mu_{f_1 \wedge f_2} = \mu_{f_1} + \mu_{f_2}$ . This is actually a linear model just like our own: written in vector notation, if  $\bar{z}$  is the binary column vector representing a category, written out as zeroes and ones, and we stack the various Gaussian locations, one per row, in a matrix  $A$ , then the location for a particular category is  $A^T \bar{z}$ .<sup>8</sup> The key differences here are that: (i) the equivalent of the predictor vector  $\bar{x}$  is latent, not observed, (ii) it is infinitely long (and so in fact we should really call it a function, not a vector), and (iii) the MLM model picks out a mixture of  $A$  matrices, whereas this IBP based model will only have one  $A$  matrix. Notice that it still gives rise to infinitely many latent categories, however, because there are infinitely many latent features. Some examples of what good models for the Inuktitut data would look like (focusing on just the non-retracted allophones) are shown in Figure 3.7.

Most notable Bayesian latent feature models are intended to extend IBP in some way. For example, the dependent IBP (Williamson, Orbanz & Ghahramani 2010) is an extension to the IBP to allow the assignment of features to observations to depend in some measure on the values of a set of observed predictors; the power-law IBP (Teh & Görür

---

<sup>8</sup>Griffiths and Ghahramani call the individual  $\mu$  values, rather than the binary values, the “feature values.” They separate out the binary feature vector as a “sparseness vector,” which we might instead call a “variable selection” vector, or a “feature selection” vector, following our terminology above. This difference with respect to the phonologist’s perspective on features is merely terminological: what phonologists call “binary feature values,” we might instead call the “feature selector values” associated with a segment; what phonologists call the “phonetic contents” of a feature,” we might instead call the “feature value,” here meaning the “value” associated with the feature. For current purposes, we will stick to the phonologist’s terminology—the IBP learns *binary feature values* plus some associated parameters, the *intrinsic content* of those features.

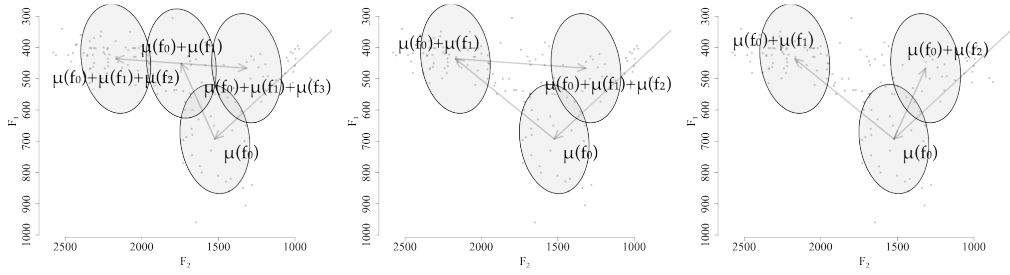


Figure 3.7: Three different ways that the clusters in a three-vowel system might be decomposed if the likelihood function in a latent feature model simply adds Gaussian means. All of the clusters here share some feature  $f_0$ , which need not be enforced (but could be). In the leftmost model we show a combination of features which is predicted to be possible but which should never be used (if the model finds a cluster assignment consistent with the right one): something like [+high][−back][−front], where we are calling the “zero” setting for a feature the − setting (although we could have called it the underspecified setting just as easily).

2009) allows the number of features used by an observation to follow power-law distributions but is otherwise the same; and, while the IBP feature assignments have the same distribution regardless of the order in which the assignments are made across observations (they are exchangeable), a general framework for handling dependencies among observations by coding relations between them in a tree structure is the phylogenetic IBP (Miller, Griffiths & Jordan 2009). The Beta process (Hjort 1990) might be seen as an exception in that it is a latent category model, and did not start from the basis of IBP (it was developed earlier); however, BP turns out to have IBP as a particular construction of a special case (Thibaux & Jordan 2007), so that all the extensions of IBP are also extensions of Beta processes. To take the Beta process as an example of a more general latent feature model: while the feature value assignments under an IBP have a single hyperparameter  $\alpha$  regulating the number of features used by an observation, under a BP prior, an additional parameter  $c$  explicitly trades off the addition of new features against old ones, similar to

the Dirichlet process prior: the probability of using a given existing feature  $k$  goes down from  $\frac{N_k}{N}$  to  $\frac{N_k}{N+c-1}$ , while the parameter of the Poisson distribution on the number of new features added is now scaled down from  $\frac{\alpha}{N}$  to  $\frac{c\alpha}{N+c-1}$ ; the IBP is the case where  $c = 1$ . Smaller values of  $c$  encourage the reuse of features, and larger values discourage the reuse of features. Once the IBP is understood, however, the general flavor of all Bayesian infinite latent feature models is fairly clear.

Of course, latent feature models need not have infinitely many features: for example, it is simple enough to add an inference for the value of  $\bar{x}$  to the MLM described above, so that the model becomes a Dirichlet process mixture of finite latent feature models. As phonological feature theories often attempt to delimit a finite universal feature alphabet, this might seem to be a key difference between phonological feature theory and the prominent current Bayesian latent feature models (and a favorable one, since finite-dimensional prior distributions are easier to work with and derive). However, there is a general uncertainty about the universality of the set of features, and so it would be imprudent to enforce this constraint too strictly; more importantly, there are deeper differences to be reconciled.

Before proceeding, it is worth making one point very clear: despite the allusions to the “presence or absence” of a feature in the current section, the binary features in a statistical latent feature model are not necessarily privative. We can think of the binary values just as easily as mapping to  $+$  and  $-$  as to present or absent. This means that we can use the exact same model to capture either full-specification binary feature models or underspecification models with privative features only; which we choose is a matter of how we interpret the feature values, and in particular how we choose a likelihood (for example, whether we choose to use a likelihood like Griffiths and Ghahramani’s, where

the 0 value for a feature implies an actual zero effect on the likelihood function).

### 3.4.3 Feature-based phonetic category models: goals for future research

In the standard IBP model, the content of a feature is drawn from some distribution which is the same across all features. When any new feature is introduced, it is always associated with a parameter value taken from the same set, drawn from the same probability distribution (like the fixed prior distribution over Gaussian location components in the Griffiths and Ghahramani model discussed above). This is at odds with the usual phonological model: if a [high] feature is used to specify some category, it is surely the case that another [high] feature will not be employed, a constraint which means something if (and only if) there is some phonetic constraint on what it means to be a [high] feature. This is possible to have even if the precise phonetic values for a feature need to be learned, and which case what is universal for a given feature is some particular distribution over possible parameter values, not a particular phonetic value. This constraint on how parameter selection works comes in two parts. Here is how we state the first part:

- (93) **FM1.** The selection of parameter values for a given feature can be separated into two steps: (i) select a distribution on parameter values; (ii) select a parameter value from that distribution. Universal phonetics approaches to feature theory claim that the distribution over distributions in step (i) is concentrated at a small number of biologically fixed points (although it would not be inconsistent with the prevailing opinion to think that some probability mass is held out for “totally learned” features).

Here is the second part:

- (94) **FM2.** Step (i) above is dependent on the other step-(i) selections in the following strong way: conditional on a distribution  $G_1$  being associated with a feature, the probability of  $G_1$  being assigned to a new feature is zero.

The other way that the IBP-like models do not align with phonology is when it comes to underspecification. In particular, the binary-valued underspecification hypothesis says that features can be omitted, not only specified as one of their two values. Conceptually, this is a matter of specifying an inference over now ternary-valued features in two steps, first, via a distribution on whether the feature will be present or absent, and then via a distribution on its binary value (in addition to specifying its phonetic content).

The Contrastivist hypothesis further asserts that there is some independent and meaningful notion of contrast which drives what gets specified lexically: in particular, certain feature specifications contain information which is not contrastive given the rest of the specification. However, as discussed above, the classical notion of “contrastiveness,” by which a feature must non-trivially partition a finite set of already-learned categories complete with binary feature specifications, presupposes a classificatory view of features, or, at least, a classificatory view of underspecification. A system learning phonetic feature specifications using a contrastiveness criterion, on the other hand, would need to use a very different notion of “contrastiveness,” one based on the degree of explanation of the phonetic data—because before learning the categories, that is all it would have available to set its criteria around.

There are therefore two very different types of models that could be built: one, the standard classificatory view, would divorce underspecification from phonetic feature specification, and build it into a second step in which the lexical representation for an observation is determined. A data point's "surface" featural specification  $z$  would determine the phonetic likelihood function, but a higher-level, "lexical" feature specification  $l$  would also need to be inferred, from which  $z$  would be predicted; the parameters of a mapping between  $l$  and  $z$  is what would need to be learned. The "surface" level does not actually to be the conventional "phonetic surface" level, (in a larger model, it might be the input to the phonological mapping from the lexical side instead), but the mapping should be able to predict features that must be present on  $z$  but which are absent on  $l$ . A single-state transducer with lexical representations subject to some simplicity-biased prior would be a good place to start.

The alternative approach would be to make the idea of "contrastiveness" translate entirely into an evaluation of the phonetic likelihood function, thus, some notion of "goodness of fit." In the IBP, the probability of setting an existing feature to 1 or 0 is determined by the number of observations with that feature set to 1 and the likelihood of the observation after setting the feature to 1/0. Consider the leftmost model, for example, in Figure 3.7: in this hypothetical set of clusters I have not drawn the clusters corresponding to  $\mu(f_0) + \mu(f_1)$  and  $\mu(f_0) + \mu(f_1)$ ; but such feature assignments ( $[-\text{high}][+\text{front}]$  and  $[-\text{high}][+\text{back}]$ ) are possible. The idea is that such feature assignments would just never be made because they would not give very high likelihood to any observations: they would not fit the data well. In a phonetic sense, there "is no contrast" between  $[+\text{front}]$  and  $[+\text{front}]$  or  $[+\text{back}]$  and  $[+\text{back}]$  for the low vowels. Rather than being interpreted



as meaning that there are no *already-formed* phones corresponding to these features, we would instead interpret “non-contrastive” as meaning there is no perceptual *evidence to support* phones corresponding to these feature combinations, that is, for low vowels, the features [front] and [back] are not *phonetically* contrastive.

In the IBP, it is not the case that there is a preference to have a feature set to zero once it is in use: sparser assignments of 1s to existing features are not “simpler” for the IBP. Rather, any time a feature is in use by at least one observation, then an additional selection of 1 or 0 will need to be made for all the other observations. However, this does mean that consequently the probability of a feature vector will be scaled down across the board, because one more probability needs to be multiplied in, and this reduction in probability is clearly an instance of the Bayesian Occam’s Razor—but with respect to *how many features are in use by at least one observation*. The relevant pair of larger and smaller model hyperparameters is a set of  $K$  previously-in-use features versus a set of  $K + 1$  features, and a BOR is yielded because the shape of the distribution over the first  $K$  feature values is the same under one of the choices for the  $K^{\text{th}}$  (in fact, both, due to the independence of feature selections for a given observation). Thus sparseness is enforced for the set of features in use taken as a whole, but once a feature is in use by at least one observation, then it is in use, and the selection must be made for all observations; the only way to get the BOR effect is to take the feature out of use entirely.

With the addition of a second inference, however, we can do the following: for each feature in use, choose whether to specify the feature; choose its feature value if it is specified. Now it is not only a smaller set of features overall that will be preferred, but also a smaller number of specified features for a given observation, because each

decision to specify a feature induces its own BOR effect on the feature specification vector for that observation as a whole. It would still not be possible to learn an outright ban on certain feature combinations because they are “redundant” in the classical contrastivist sense—this model still ultimately relies on the likelihood function to assess “contrast” and unusual observations still might give rise to unusual feature combinations—but it would at least give rise to the possibility of underspecification of binary feature values, and there would be a preference for underspecification.

To sum this up, the two possible versions of contrastive underspecification are as follows:

- (95) **FM3a.** For each observation, the feature specification must be converted to a lexical representation via some mapping that must be learned. Underspecification can arise if the mapping enforces sparseness on the number of features.
- (96) **FM3b.** The specification of feature  $k$  for observation  $i$ ,  $z_{ik}$ , is a two-part choice: first between 0 (underspecified) and 1 (specified), and then between  $+$  and  $-$ , conditional on the first choice coming out as 1.

The first choice relies on a separation between lexical and non-lexical featural representations, via a mapping which sets out what are and are not possible (or likely) lexical representations; the mapping needs to predict non-lexical from lexical feature specifications. Enforcing sparseness of lexical representations can therefore penalize lexical representations which are equally predictive but more complex (like full specification versus underspecification). This is a version of the classical contrastivist idea: features will be underspecified if their binary values can be predicted from other values. The second

choice does not make such a separation. This is a model where the choice to specify or underspecify relies only on how well the observed *phonetic* values can be predicted from one specification or the other. It relies only on what might be called a notion of “phonetic contrast,” and it does so no more than any other phonetic category learning model we have talked about.

Finally, we must discuss the likelihood function. Above, we gave examples of what phonetic features might be learned under a Griffiths and Ghahramani-type linear Gaussian model. These models did not look very much like the usual feature models in phonology, because a feature value could either be 1 or 0, and, if it was 0, it induced a zero change in the likelihood. There was no way to have a single feature give rise to two different non-zero effects. However, recall our change in coding in the gender model above: the same set of induced categories can be learned with a different representation if we code the feature values as 1 and  $-1$  rather than 1 and 0. See Figure 3.8 for an example of what this might look like. It is much more like what we would conventionally expect phonological features to do for vowels. Strictly speaking, it is not necessary to change to an underspecification model in order to use this coding, (a standard IBP model would work), but that is necessary in order to get a model to come out looking like Figure 3.8: the low vowel category /a/ is underspecified for backness.

A more important difficulty in specifying the likelihood function is the problem of the covariance. The problem does not arise for Griffiths and Ghahramani, because their features have fixed covariance. Recall our discussion above, in which we said that we would deal with the problem of  $\oplus$  needing to apply to full category map “phonetic representations,” which somewhat oddly, implied that we could sometimes add the shapes of

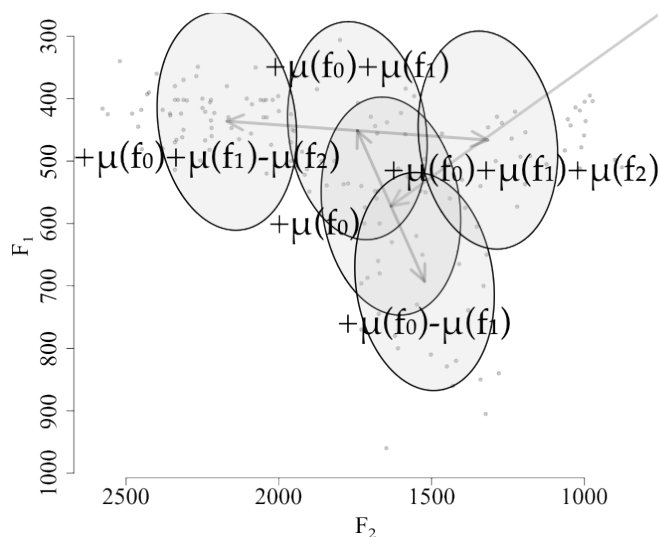


Figure 3.8: One way that the clusters in a three-vowel system might be decomposed if the likelihood function in a latent feature model with feature values  $+1$  and  $-1$  adds Gaussian means. All of the clusters here share some feature  $f_0$ , which need not be enforced (but could be). If points are to be assigned assigned to the cluster with center at  $+\mu(f_0) - \mu(f_1)$ , then this model is only possible if there is also a zero value—that is, if it is an underspecification model. Those points (the vowel /a/) will need to be underspecified for the feature  $f_2$ , the backness feature. Under the model discussed in the text, it will still be possible to assign tokens the feature values  $f_0 = 1, f_1 = -1, f_2 = 1$  and  $f_0 = 1, f_1 = -1, f_2 = -1$ , corresponding to the locations  $+\mu(f_0) - \mu(f_1) + \mu(f_2)$  and  $+\mu(f_0) - \mu(f_1) - \mu(f_2)$ , but we would expect these points to have implausibly low likelihood.

phonetic categories together, by saying that it simply never arose that we would need to do that. Now, we need to do something almost exactly like that: we need to build categories out of smaller pieces, and the linear Gaussian model says that we add the pieces together (with  $+$ ). It is almost certainly not the case that we could coherently account for all the differences in the shapes of (perceptual) phonetic categories that arise using addition of covariance matrices. There are a number of ways we could get around this while keeping the Gaussian assumption, (we could hold the orientation of the Gaussian fixed

and hope that we could get away with adding and then squaring to obtain the eigenvalues of the covariance matrix), but all would add some complications to inference. A proper exploration of how the featural composition of a phonetic category determines its shape, if at all, is in order; it would be interesting to take the idea seriously that the addition of features is by  $\oplus$  and thus affects the shape in the same way as phonetic transforms do.

I will leave an implementation for future research, but it is worth reviewing some of the technical issues here for the interested reader: FM1 makes the likelihood more difficult to compute—but not impossible, particularly if it is possible to integrate out the parameter values that go into the likelihood functions; FM1 places unresolved empirical demands on the construction of the model—but the assumption could be dropped; the research strategy would be to assume that features are totally learned until modelling results deem it necessary to add a bias, and the research program would begin by evaluating how well modelling results align with the linguistic typology of features; if there is truly no held-out probability for learned features, then FM1 makes the feature model finite, which is easy to work with; FM2 makes the distribution on the sequence of parameter values nonindependent, but it could be dropped if UG were such that choosing the same feature distribution twice would be unlikely to yield high likelihoods (or simply by dropping FM1 altogether, as discussed); FM3a requires additional machinery be added to a standard IBP model, but there is nothing particularly novel about that machinery (including a probability of adding particular features to  $z$  conditional on the makeup of  $l$ ); FM3b constitutes a fairly small change to the IBP. Finally, and perhaps most importantly, we need to work out a proper model by which combination of features translates into combination of parameters, keeping in mind our concerns about specifying the covariance.

### 3.5 Further issues

Consider again the Inuktitut vowel data, shown in Figure 3.9. This data shows allophones with different covariances. Under the real-space hypothesis, we can interpret the size and shape of these distributions directly as representations of components of a phonetic map (or at least what must go into it and come out correctly identified).

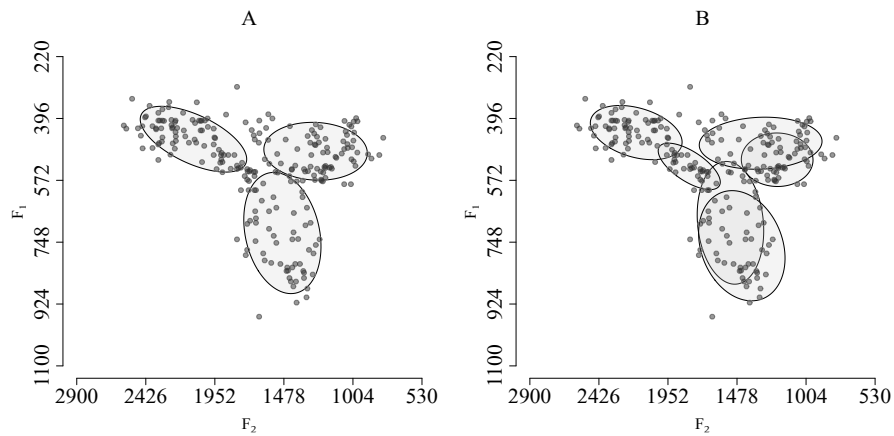


Figure 3.9: Second and first formant values for Inuktitut vowel tokens (repeated from Figure 3.3).

This does not comport with how we have implemented the LPT. There are at least three possibilities:

*Additions to covariance* It is possible to make whatever adjustments we need to the scale parameter of a Gaussian of using addition; more generally, we can also adjust whatever representation of a category we have using addition, including the scale (on some other representation the location and scale might not be independent).

*Changes to  $\oplus$*  The way that  $\oplus$  works is not simple addition; it only looks like simple addition if we consider the location of categories.

*Scaling of dimensions* The perceptual dimensions are scaled differently from the acoustic dimensions. The categories do have the same covariance matrices in perceptual space.

The idea behind the last suggestion is that the vowel space might be on a different scale, and, if we simply changed the scale on the axes, we would get all the covariances looking like they were the same shape across allophones. (There are, of course, various psychophysical scales—bark, mel, etc—which are presumably more like the low-level auditory input than Hertz, but these are fairly close to linear for much of the vowel space, and at any rate do not impact the appearance of this data much.) The problem with this is that the allophones, taken as a whole, seem to be rescaled in a way that does not depend only on their location in the vowel space, but also on the fact that they are allophones: in addition to being shifted downward and back, they are scaled down in their variances. If the scale being different were the thing responsible, we would expect the scales to be the same between allophones of a single phoneme, given how close they are to each other in formant space.

The first two possibilities are very difficult to distinguish, as it will generally be possible to change our representation of the shape of the category; the details will depend on both the evaluation measure and the way scale effects track  $\boxplus$ .

I will leave the empirical details to be worked out; but there are several interesting patterns that can be found in the retraction data, which may prove relevant. First, the retracted vowels have smaller variance in both dimensions than the non-retracted vowels, and the retracted low vowel also has a smaller variance  $F_1$ , although the variance increases

in the  $F_2$  dimension; this suggests there might be something systematic about the change in scale, either tracking the allophonic/underlying status of the category, or the degree to which it has been retracted. The pattern is imperfect, and we will see in a moment a case where the changes in scale are somewhat more arbitrary; but we will also see a suggestion of a pattern like this again in Chapter 5.

The vowels are also closer together when they are retracted. This is almost certainly due to physical factors: limitations on the position of the tongue demand that the low vowel not be shifted as far as the two high vowels. This should influence our thinking on how  $\oplus$  works: as we reach the edge of the vowel space, the sample covariances must necessarily become more compressed. Whether the perceptual maps reflect the sample covariances in this way is a separate question, but if they do, then  $\oplus$  needs to operate in a way that takes this into account.

We might instead be led to think that the locations follow from the absolute position in phonetic space to some degree: perhaps the three are specified as equal transforms, but the computation of  $\oplus$  corrects for the limits on possible vowels. There are some facts, however, that suggest strongly that this is not the explanation for the compression of the vowel space. Consider the example shown in the leftmost panel in Figure 3.10 of retraction in Kalaallisut, which is another Inuit language, the official language of Greenland; it has the same vowel inventory and the same rule as Inuktitut.

The retraction in Kalaallisut appears to be more pronounced than in our Inuktitut data. The variances show a different pattern than in Inuktitut, as they are not clearly scaled down in the retracted vowels in the way they are in Inuktitut. There is some reduction in the variance in the  $F_2$  dimension, but this could just as easily be attributable to the effect of



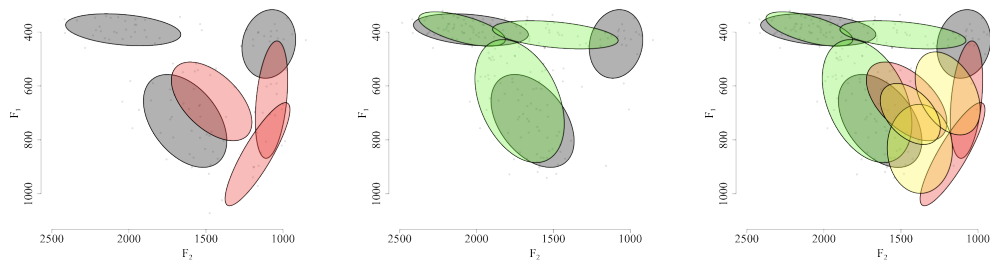


Figure 3.10: Vowel categories from Kalaallisut (see below for a description of the corpus). The vowel ellipses in the left panel are unretreated (grey) and retracted due to the influence of a following uvular consonant (red); in the middle panel, not fronted (grey) and fronted due to the influence of a preceding coronal consonant (green); in the right panel, unretreated and unfronted (grey; in fact, all the grey ellipses in all three panels are unretreated and unfronted) and both retracted and fronted (yellow).

reaching the back of the vowel space. Now, Kalaallisut also has an allophonic rule which fronts vowels slightly, particularly /u/, after coronals. This is shown in the middle panel. Here, again, the vowels are closer together than in their default pronunciations, but this, too, could be attributable to reaching a corner of the vowel space, and it could literally be the effect of a grammatically active constraint on possible outputs, limiting the ultimate locations of the targets—rather than a rule which happens to respect this constraint. This might be our explanation if we considered the two processes separately; but it cannot be correct. The right panel shows what happens when both processes apply: the vowels that have been both retracted and fronted are still tightly clustered together, even though they are now in the middle of the vowel space. The action of the two transforms is such that the vowel space winds up compressed, although there is no physical reason for it to be so any longer.

We may also be able to use this case to shed light on the operation of  $\oplus$  per se. For the retracted vowels, the difference between fronted and unfronted /u/—where fronting

is most pronounced—is far less than in the unretracted /u/, a difference of about 122 Hz in  $F_1$ , as opposed to 395 Hz for unretracted /u/. This suggests that  $\oplus$  is not simply doing addition of the (rescaled) acoustic values. It should not be taken as relating to associativity or commutativity: we have been given no independent evidence about different ways of combining the effects of transforms. Rather, we have been shown *one* way—the actual way—that phonetic transforms combine, and we have seen that it deviates from ordinary addition. One approach to handling this might be that the phonetic representations are all relative, not absolute, and they therefore combine by something more like multiplication—but how to apply this so that it allows us to state the full range of transforms, and at the same time state the categories themselves, is not totally clear.

Another issue raised by the application of this model to talker variability—or rather, by the presentation of a concrete model of talker variability perception—is the distinction between *I-language* and *E-language* (Chomsky 1986). An I-language is simply whatever (“internal”) state of a particular human’s brain permits it to process, produce, and assign meaning to language utterances—the mental grammar, but the term is set up to contrast against an *E-language*, which is an (“external”) set of utterances, or utterances paired with referents, or situations, or whatever other collective externalization of language one could find to label “a language.” As Chomsky points out, it is much easier to say precisely what constitutes an I-language than what constitutes an E-language, and, in spite of some researchers who take the study of language to be the study of E-languages, it seems more reasonable to study I-languages, not only because the study of mind is interesting, but also because we ought ultimately be able to derive whatever properties an E-language happens to have from properties of the minds that speak it.

One nice consequence of this is that it deproblematizes the question of what constitutes a “language,” familiar from comparative linguistics: why are Bosnian/Serbian/Croatian/Montenegrin different “languages” when their standard forms are in fact basically the same apart from the reflex of Proto-Slavic yat, while the three sub-“dialects” of Croatian (Shtokavian, Kajkavian, and Chakavian) are not mutually intelligible? The answer is that it is beside the point; the terms “language” and “dialect” are put to inconsistent use depending on the circumstances, and the desperate need for a clear distinction is motivated only by the misconception that E-languages are natural kinds. The I-language perspective makes it all right if they are not, and if, say, E-languages are only derivative properties of interacting collections of I-languages, which is to say, individual people’s grammars which have slightly different properties from one person to the next. Grammars may be more or less similar, but there need not be any meaningful points at which two grammars change from being idiolect-different to dialect-different to language-different.

Nevertheless, taking the speaker model to its logical extreme, in which each speaker has a separate sub-model, there is a question that is raised: why does the learner construct a common model for all these speakers in the first place? That is, why are the speaker-dependent phonemes not simply separate phonemes? After all, if everyone around you has a different grammar, why not simply take this at face value? Why assume that they speak the same “language”? There are reasons, of course: simplicity biases militate against simply adding new lexical items, phonemes, and so on, just because they were spoken by someone different; having a noise model, without which even within-speaker variability would be impossible to process, allows slightly different pronunciations to be assimilated across speakers as well; and the apparatus must exist to handle allophony within a sin-

gle speaker, and so effects of talker-level variables certainly *can* be partialled out in the phonetic grammar. Nevertheless, the world did not have to be this way: it could have been the case that I-languages were devices which the brain constructed specially to operate on input from one single other I-language, but it is not the case: although they can contain adaptations to handle individual-level variability, I-languages are not (only) models of individual I-languages. In the speaker model, I set the scalars coding male/female, young/old to be  $-1$  and  $1$ , which meant that the common, neutral content of a category did indeed have a cognitive status (it is the intercept in the  $A$  matrix); even if this had not been the case, the common category label would have been shared across different levels of the talker variables despite the fact that the phonetic interpretation of the category at each of those levels differs. Thus, in one way or another, properties common to speakers, and not only to individuals, have a cognitive status.

This is not surprising, of course, nor does the fact that the grammar gives some internal status to a common “language” mean that that “language” must align with constructs like “English,” “Dutch,” or “Chinese.” (Presumably, however, there is some status given to arbitrarily constructed language/dialect-level indicators that allow us to make use of the information that a speaker used the word *yall* or *vous-autres* to switch us, categorically, to a system that will make sure we say *coke* or *liqueur douce* rather than *pop*, as we might normally be inclined to say—and which will, more importantly, give us some general expectations about that speaker’s phonology, grammar, vocabulary, and so on.) Nevertheless, it does mean that “I-language” should not be the lens through which we view everything.

Finally, the discussion of a model in which features rather than atomic categories are

learned raises the question of what, then, it would mean to be an allophonic process at all for the purposes of our conjecture: that allophonic processes—those processes that do not output existing lexically contrastive categories—are phonetic and gradient. What would it mean if there was no primitive notion of “category”? We could modify the criterion in one of two ways: we could either say that the output of a process was a strict allophone if it generated a novel combination of (possibly elsewhere contrastive) features; or only if it made use of a nowhere-contrastive feature. This latter is dangerous, however, as it implies that allophonic processes simply add feature–value pairs, (or at least can), which is at odds with the way we described phonetic transforms as working. The problem is that the notion of “non-contrastive category” in the definition of a strict allophone was really intended to be a descriptive, or even an external, description of a phenomenon, for the purposes of attaching an explanation in the form of phonetic transforms; it does not need to be replaced when we move from an atomic representation of lexical segments—which are *internal*—to a feature-based one. I discuss these issues at greater length in Chapters 4 and 5.

### 3.6 Summary

In this chapter, I have highlighted the existence of context-dependent phonetic grammar, and put forward a new hypothesis about it, namely, that context-dependent phonetic transforms are linear and additive. I have used this to show not only how we can learn the phonetic contents of allophonic processes, and how this can improve our ability to learn phonetic categories, but also to show how we can learn the environments in which

they occur. I have also gone over the basic statistical principles by which we can set up phonetic category models, and shown what would be involved in making such a phonetic category model reflect the idea from phonology that categories are inherently decomposed into features (as well as some of the barriers to doing that). I have presented some phonetic data bearing on the way that the context-dependent phonetic transformations work. Finally, I have raised a further conjecture: all allophonic processes are phonetic. The rest of this dissertation investigates this idea further.

## Chapter 4: Phonetic transforms I: The cognitive architecture

The spirit talks in spectrums

He talks mother earth to father sky

—Joni Mitchell, “Don Juan’s Reckless Daughter”

### 4.1 The phonetic surface

Chapter 3 introduced the *linear phonetic transform* hypothesis (LPT) and then put forward the conjecture that allophony is due to the application of phonetic processes. Before we begin any discussion of the implications of this, we need to say what we mean by “allophony.” We are referring to the existence of a certain type of relation between phonetic categories, which is to say a certain type of relation between equivalence classes over phonetic tokens.

What kind of equivalence class? Take it as given that we can sort out which segments belong to which categories, including the categories we would like to identify as allophones, not as learners but as analysts: that is, suppose we are given a descriptive oracle which can turn a speech signal into a sequence of what are ordinarily called *phones* (unlike the mixture of Gaussians procedure discussed above in the first set of Inuktitut experiments, which was *not* good enough to deliver such an oracle). Although the reader

may have already gleaned that our theory implies that nothing even approximating this oracle is part of phonological cognition, in order to have a coherent definition of what constitutes an allophone, phones must exist descriptively. Whether such an oracle would be based on acoustic statistics, or some kind of measurement of the muscle movement involved in production, or whether it would actually be impossible to build, talk about allophones, descriptively, is talk about the output of such an oracle; that is all we are saying here.

Given an oracle, there are two related patterns I will be folding under the term “allophony.” First, the very shallow property of *complementary distribution*. Complementary distribution simply says that a pair of oracle categories never occur in exactly the same segmental context (which we need to take to mean “in the same  $n$ -phone sequence, for some reasonable  $n$ , often 3). If we take any transcription from the oracle, take two categories and replace all instances of both with question marks, then, if they are in complementary distribution, the remainder of the transcription is always sufficient information to restore the question marks, and conversely (up to a minimal  $n$ ). Now, there are many things that get called allophony that do not show complementary distribution, and vice versa:

*Non-surface allophony* “Obscured” complementary distribution is often still considered allophony. Take the words *write* and *ride*. With respect to the pronunciation of the second one, [rajɪ], the pronunciation of the first is relatively short, central, and slightly raised in many Northern dialects of North American English: [rɪt̚]. If an oracle could distinguish [j] and [aj], they would be in complementary distribution in



many cases like this. The key difference is the voicing on the following segment. But North American English also pronounces both *writer* and *rider* with the flap [ɾ] as in *water* [wəɾ] in place of both the coronal stops [t] and [d] ([ɾ] is in complementary distribution with both of those). As a result, *writer* is pronounced [ɾjər] and *rider* [rajər]. So, in fact, the oracle would tell us that, looking at this larger sample of the language, [j] and [aj] are *not* actually in complementary distribution. Still, they are “allophones” under many descriptions, and especially if the phonological grammar works in multiple steps, of which a conversion of [t] and [d] to [ɾ] is just one. At the level of representation before this conversion happens, [j] and [aj] do stand in complementary distribution. Given the oracle, they do not stand in complementary distribution, but they fall into the same class as things that do if we assume a bit about the grammatical system that underlies this complementary distribution.

*Mismatching pairs* In Spanish, [b] is in complementary distribution with [β] and [d] is in complementary distribution with [ð]. However, as [β] and [d] have the exact same restricted distribution, it is also true that [b] is in complementary distribution with [ð] and [d] is in complementary distribution with [β]. Nevertheless, [b] and [ð] are not considered allophones, and neither are [d] and [β]. Again, this is because “allophony” refers to one of a number of descriptive generalizations intended to feed a theoretical account of what is going on, and, in particular, the distributional relation between [b] and [β] is thought to be the result of the grammar turning [b] into [β], but the distributional relation between [b] and [ð] is not thought to be the result of turning [b] into [ð] (similarly for the other case). Again, descriptively, if all we presuppose is

the transcription, then all four pairs stand in complementary distribution, but they do not all fall into the same class once we add this particular theoretically-committed rider about what is going on cognitively.

*Crazy pairs* In English, [ŋ] and [h] are in complementary distribution, because [ŋ] only occurs syllable-finally and [h] never does, but no one would say they are allophones for the same reason that [b] and [ð] are not allophones. This is only different from the mismatching pairs case in that the two are not even supposed to be indirectly related; it is a complete coincidence that the two have complementary distributions, and no one attempts to claim that there is any common cause in the form of a process giving rise to either of these two positional restrictions.

In short: *allophony* is some causal account for (certain) cases of *complementary distribution*, which, given an oracle that allows us to talk about phones, (“oracle categories” henceforth) is itself just a descriptive generalization. Although both may go beyond a finite corpus, allophony is something reasonable people could disagree about, and complementary distribution is not (again, assuming they had access to the transcription oracle).

In addition to complementary distribution, I will also relate phonetic transforms to *strict allophony* (see Chapter 1). Now, this idea presupposes more than an oracle: the idea is that there are oracle categories that appear in the transcription but have no corresponding lexical category. Thus it assumes an analysis of the lexicon. The first important idea about this will be that phonetic transforms, regardless of whether they are patterns that would traditionally fall under “allophony,” (in Chapter 5 we will see some interesting cases where they are not), are virtually assured to give rise to an analysis with almost exactly that shape:

if we had an oracle, we would find categories that emerge from the phonetic grammar but are not coded lexically. (As we will see, one difference is that the categories are not given a categorical phonological representation *anywhere*.) The other term for the allophonic processes in analyses like this is *non-structure-preserving*: they introduce information that cannot (or at least is not) coded in the lexicon. The second part of this chapter will discuss some of the known descriptive generalizations about non-structure-preserving processes, and show how the current analysis hangs on to those.

Now, barring the distributional pattern being obscured by other processes, such situations should give rise to surface distributions that are complementary; and this idea of surface-only categories is hardly theory-specific. However, as referenced in passing in Chapter 1, depending on the theory, that may or may not be the mechanism accounting for complementary distribution patterns. We will see presently that analyses in Optimality Theory usually take a slightly different approach.

The rest of this chapter will thus proceed like this: first, take it for granted that phonetic transforms account for complementary distribution, and note that such an analysis is a kind of non-structure-preservation, or strict allophony analysis of complementary distribution (and of course other cases—to be discussed in Chapter 5). The first part of the chapter then discusses a major consequence of this, which is that there is no surface representation in the conventional sense. This puts the analysis at odds with the analysis of many of these patterns in Optimality Theory as “surface phonotactics”; and it puts it at odds, albeit less sharply, with the classical analysis as categorical rules. I will then explain why the architecture is the way it is, and not otherwise, in terms of the special nature of phonetic non-structure-preservation; this will be seen to be related to a known empiri-

cal pattern I call the “lateness of allophony.” Finally, previewing Chapter 5, I will start to sketch the reason that not only the analysis, but the *theory*, is at odds with the conventional surface representation—in other words, why the architecture should lead us to treat complementary distribution patterns as non-structure-preserving phonetic transforms. The key will be Bayes’ Rule.

#### 4.1.1 Background: Surface representations

The way we have talked about learning phonological categories up to now, it has been as if there were a finite set of lexical categories: possible segments that can be stored as part of a lexical item. We made some brief reference to the fact that our statistical models do not really comport with the assumption of finite inventories, but it is a standard assumption in linguistics (and, at any rate, they *do* imply that inventories are enumerable). To understand the idea of a finite lexical inventory, we need to first ask what it means cognitively. The idea is simply that there is *some* restriction such that every stored lexical item will always be a string over some finite alphabet *LI*, the lexical inventory. They did not always agree on how these restrictions were to be specified, but phonologists thinking cognitively thought this way for the better part of the twentieth century: Jakobson 1941 tracks the child’s development of phonemes, (the “sounds that are used to distinguish the meanings of words,” 29), and makes it clear implicitly that, at each stage, including the adult state, the set of possible phonemes remains finite (“while the succession of phonological acquisitions in child language appears to be stable ..., [t]here are children who ... still have not completely mastered their phonemic system at school age,” 46); much of

this follows from the understanding that lexical inventories are made up of specifications of binary phonetic features, and the universal set of possible phonetic features is finite (“In the succeeding pages we shall list the individual features that together represent the phonetic capabilities of man”: Chomsky & Halle 1968, 299). However, the usual view goes beyond just this, and asserts that the lexical inventory of a language is a subset of this finite universal lexical inventory; in other words, there is a hard, language-specific restriction on what can and cannot be coded in the lexicon. This view is to be found in most work bearing on the question of lexical inventories through to the beginning of Optimality Theory. To take some notable examples: Halle 1959 derives this from a general economy condition, “Condition (5): In phonological representations the number of specified features is consistently reduced to a minimum”; similarly Chomsky & Halle 1968: “Languages differ with respect to the sounds they use and the sound sequences they permit in words. Thus each language places certain conditions on the form of phonetic matrices and hence on the configurations of pluses and minuses ... that may appear as entries in the classificatory matrices of the lexicon” (381); Aronoff 1974 constrains a class of “allomorphy” rules by enforcing an early version of a structure-preservation condition, “that they cannot introduce segments which are not otherwise motivated as underlying phonological segments of the language,” implying, given the finiteness of the universal inventory, that the lexical inventory is also finite; Kean 1975, when she writes, “Of the set of possible segments characterized by the distinctive features, it is evident that some are present in nearly every language, with others only occasionally occurring,” (6–7), can also be seen as implying such hard restrictions, depending on what is meant by “present”; Archangeli 1984 derives restrictions on lexically possible segments from an underspecification condition:

“In a language’s underlying representations only the features that are distinctive in that language, that is, features which actually are necessary to distinguish two sounds, have values specified,” (43); and Avery & Rice 1989 introduce the Node Activation Condition — “If a secondary content node is the sole distinguishing feature between two segments, then the primary feature is activated for the segments distinguished. Active nodes must be present in underlying representation,” (183), to be held in tension with the Universal Markedness Theory, which “supplies minimal structure, ensuring that unmarked values are absent in underlying representation while marked values are present,” and “can be overridden only if the NAC requires additional structure,” (184). Since specifying an unmarked value would then immediately allow for a host of new segments to be represented, this implies that lexicons are forced into not only not representing any of these segments, but not *being able* to represent them, if it is not needed to distinguish (presumably) any pair of lexical items. In all of this work, there is some mechanism that is part of language-specific or language-universal phonological cognition that limits the possible segments that can be stored in lexical memory to some finite language-specific set, a *lexical inventory*.

This is different from a *surface inventory*. That is, it is different from a restriction on what segments are phonetically possible. What might such a thing mean? At the external level of realized gestures and auditory signals, the notion of the inventory, in the sense of hard, language specific restrictions on what is possible, is not really very useful; any look at ultrasound or auditory data (or the individual acoustic points on the vowel plots found throughout Chapter 3) will confirm that the set of possible phonetic realizations is not only infinite but extremely broad. While it might be possible to rule out a few things entirely, like English speakers pronouncing clicks, even this would be quite tenuous. This

is not what is usually meant by a surface inventory. Moving up the chain of cognition: there is substantial ambiguity (not quite disagreement) in the literature as to the role of binary features in the cognitive systems that are closest to these external states, the motor and perception systems. Still, it is usually understood that there is some level of very fine-grained representation at least coming close to capturing the detail in the continuous externalization. Thus it is not at this “gradient phonetic level,” either, that the notion of a surface inventory is appropriate.

Rather, consider the level we have been dealing with in Chapter 3: a set of phonetic categories, each specified as some parameter values (see Chapter 3), and these parameter values each in turn specifies a different phonetic recognition model. Now, these phonetic recognition models *deal* in graded data, coming from this “gradient phonetic level,” but the models themselves form an enumerable set, regardless of whether they form a mixture of Gaussians or a mixture of linear models. More importantly, that set is a subset of the whole set of possible phonetic recognition models. This is a crucial idea that we will use below: whatever a language might specify in terms of how binary features are realized in the gradient phonetics, **as long there are fewer possible categories than possible gradient phonetic realizations for them, then in a sense we can say there is a “surface inventory,” an inventory of phonetically interpretable segments (surface categories).** A *surface representation*, then, is a representation consisting exclusively of elements of the surface inventory, the output of a stage of phonological grammar (seen as the mapping *from* lexicon *to* phonetics) that deals exclusively in changes from lexical categories to others, or to (exclusively) surface categories.

As we alluded to in Chapter 3, the phonological analysis that took place up to the

early twentieth century was done without access to detailed phonetic measurements, and, when consideration was given to phonetic detail, the detail was generally given a categorical description (for example, all of the phonetic transcription systems invented for spelling reforms or for transcribing unwritten languages throughout the nineteenth century, including the IPA, were finite alphabets, despite often capturing relatively fine details—see Kemp 1994; and Sapir, while often giving quite vivid phonetic description, nevertheless generally only did this in order to provide a detailed articulatory explanation for a given sound). Sapir, and to a greater extent Halle, in *The Sound Pattern of Russian*, and to a still greater extent Chomsky and Halle, constructed phonological grammars as collections of rules changing lexical representations that hewed precisely to a set of already laid-out “possible segments,” including the rules that handled allophony. Of course, there was no principled reason for this: Chomsky and Halle’s idea that role of the phonological grammar was, uncontroversially even now, a system that would start from binary features and “gradually convert these specifications to integers,” (65), a statement which tolerates rules dealing in these integer values—but they did not do this other than for stress. So, given that complementary distribution and related patterns were explained as cases of strict allophony, and given that this allophony was handled by changing representations from one category to another, with all the rules applied to a lexical representation, one would therefore derive a surface representation with all the allophones marked as separate categories. We will call such a representation an *SC-representation*. Crucially, such an analysis implies that the SC-representation has a cognitive status—that is to say, it makes it part of a structure induced by the model phonological system which we could, and, under the strongest interpretation, would, demand be shared by perception, production, and/or



learning, under some homomorphism.

Certain theories place explicit demands on SC-representations and are very difficult to reformulate without them. Key to *natural generative phonology* (Hooper 1976) is the *True Generalization Condition*, which requires that “all rules express transparent surface generalizations, generalizations that are true for all surface forms,” (13). This could be used to prohibit, say, the analysis of *writer* and *rider* described above where there is a rule changing [aj] to [j] in an environment that does not exist in the SC-representation—before voiceless stops, (where the [aj] has primary stress or the following vowel is unstressed), whereas the flap [ɾ] is not voiceless. The generalization is true at one step in the operation of the proposed phonological grammar, (before flapping), but not in the SC-representation. Now, if there is no SC-representation in which [aj] and [j] would appear, then there is no way that the NGC can even be evaluated as stated for this particular pattern, because there is no representation supporting a generalization about the oracle category [j]. It only makes sense to talk about comparing “the generalization” expressed by a rule to the SC-representation if the SC-representation exists and the rule is a categorical rule. The same holds for the voicing of the flap: if the flap is only present in a gradient representation, then there is no way for the NGC to be evaluated for the pattern in question. Again, the SC-representation is different from just any “surface representation.” The SC-representation is one in which allophones, whatever is to be said about them, are distinguished as segmental categories.

All of this applies equally to any theory in which the grammar or constraints on the grammar crucially refer to surface representations: if there is no SC-representation, then the scope of these generalizations needs to be carefully re-examined. Crucially output-

referring constraints are precisely the direction that phonological theory took starting in the 1980s (McCarthy 1986, Paradis 1988) and culminating in the development of Optimality Theory in the early 1990s, (Prince & Smolensky 2004), which is a theory of grammar in which all the free parameters are statements that make crucial reference to outputs. These outputs are nearly always understood to be categorical (see below for some discussion of exceptions), and at least some of the grammatical statements made in these theories are crucially SC-representations (namely, those that make reference to allophones)

Regardless of whether the phonetic allophony proposal turns out to be correct, a re-examination of all these cases needs to be made: there is no argument for an SC-representation to be found anywhere in the literature. It follows in a theory in which the patterns that are complementary-distribution-like are all explained by appeal to categorical changes made to underlying representations that do not code certain non-contrastive segments; but there was never any requirement that we do so (and, in fact, Chomsky and Halle reference Sledd 1966 in the context of gradient rules; he describes Southern American English dialects using context-dependent gradient rules, albeit stated in words).

Now, in the usual Optimality-Theoretic formulation, furthermore, complementary distribution is in fact often *not* accounted for by changes made to underlying representations via the grammar; it is not always treated as allophony in the non-structure-preserving process sense. Rather, since the theory turns around the idea that the learned aspects of the grammar only pick out different ways of constraining outputs, or lexicon–output mappings, and not lexical information per se, there is no hard lexical inventory in standard Optimality Theory. The result is that there is no restriction on which of the SC categories should be stored, given that an SC category is part of a complementary distribution pat-

tern, and so a general principle (*Lexicon Optimization*) pushes the learner to store the observed surface category. The only case where this really would not work is understood to be active morphological alternation (*titirauti–titirauti+qaq+tunga*), because that other information constrains how a morpheme should be stored. Cases of complementary distribution without morphological alternations are usually treated as phonotactic knowledge, constraints on legitimate sequences in surface representations; and these are obviously stated over SC-representations. This is the usual analysis even when the very same segments also some times participate in alternations: in some cases, a surface constraint will lead to an actual alternation, but in other cases the same surface constraint will just express a generalization about possible sequences of segments morpheme-internally, a static generalization about lexical items stored faithfully with the surface SC categories.

This reduced burden for the phonological grammar making changes in explaining complementary distribution actually undermines what little motivation there was for the SC-representation in the first place (*if* complementary distribution is due to grammatical processes, *and* the only grammatical processes are categorical—even though no one ever said this—*then* there must be an SC-representation). The use of SC-representations is an unexamined assumption that would change the analysis very much if it were dropped. Prince and Smolensky’s discussion of the fact that additional mechanisms of “lexicon optimization” along the lines of Chomsky and Halle’s symbol-counting evaluation measure are needed to even get the learner to unify two pronunciations of a single morpheme in one lexical entry could also have led theoreticians to question the value of SC-representations, but, so far as I know, it was never vigorously pursued.

There is empirical work to be done in pursuit of the question of whether SC-representations

exist as well. Speech perception seems to generally treat allophones differently (see Chapter 3 for some discussion). These have yet to be properly explained, and they are especially puzzling under the theory that complementary distribution is just like any other phonotactic pattern, because the behavioral pattern is often just the opposite of what would be expected intuitively: listeners are *less* sensitive to the distributionally wrong allophone being used, in spite of the fact that complementary distribution is the strongest possible case of distributional regularity, which ought to make illicit sequences extra salient. Another problem for SC-representations, whatever the analysis of complementary distribution and its ilk, would be if perception under allophony was crucially determined by gradient properties of the signal in ways that comparable contrastive category perception is not.

#### 4.1.2 Status of surface representations under a phonetic transform hypothesis

In this section, I outline the phonetic transform architecture, and say why accounting for complementary distribution patterns as the result of phonetic transforms leads to our throwing away the SC-representation in favor of a more abstract surface representation, which I will call the *AC-representation*. Figure 4.1 is what the architecture of phonological grammar looks like under the current view, compared to what it looks like under the conventional view.

The current architecture, shown on the left in Figure 4.1, is one in which, like in conventional views, lexical memory is a collection of strings, and the role of the cat-

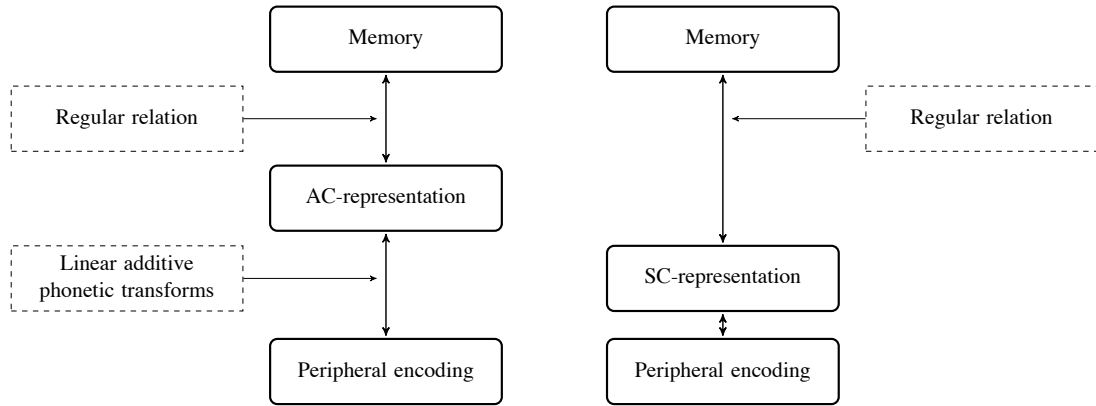


Figure 4.1: Current (left) and conventional (right) phonological architectures.

egorical phonological grammar is to map from a combination of these strings to some output string (or vice versa). It is well known (Johnson 1972, Kaplan & Kay 1994) that no known phonological pattern exceeds the computational capacity of a finite state device, which is to say that, in general, the phonological mapping is some regular relation. This is a descriptive, not a theoretical statement: even if some theory of phonological grammar can express non-regular relations, actual phonological grammars are not. As in the conventional view, too, the surface representation must be translatable into a gradient representation interpretable by a peripheral (sensory or motor) system.

To say a bit more about how this works here: the AC-representation specifies a sequence of segmental categories  $z_1 \dots z_k$ ; selecting a  $z_i$  induces a context representation  $\bar{x}_i$  constructed as a function of  $z_1 \dots z_{i-1} z_{i+1} \dots z_k$ . We have not said exactly how this works, but a simple way to understand it would be a concatenation of all the feature–value pairs, as in Chomsky & Halle 1968, Chapter 8. Each  $z_i$  also provides a context-independent category representation  $r_i$ , and a function  $t \circ T_i$  associated with category  $z_i$  that applies to  $\bar{x}_i$ —in the real-valued version presented in Chapter 3, we multiply  $\bar{x}_i$ , a vector of 0/1 or

— 1/1 feature specifications, by a matrix of independent context effects—and return a “net contextual shift,”  $\underline{s}_i$ . The final gradient peripheral encoding is then obtained as  $\underline{p}_i = \underline{r}_i \oplus \underline{s}_i$ . Crucially, although nothing about the architecture actually commits us to seeing  $t \circ T_i$  as literally treating the context elements as scalar multipliers for some context effects, in the way that the real-valued Chapter 3 version does, we *are* committed to  $t \circ T_i$  being a *linear additive* function of  $\bar{x}_i$ . The context  $\bar{x}_i$ , whatever its content is exactly, can be viewed as a combination of independent formally possible contexts,  $\bar{x}_i = \bar{a}_1 \boxplus \cdots \boxplus \bar{a}_m$ , and  $\underline{s}_i = t \circ T_i(\bar{a}_1) \oplus \cdots \oplus t \circ T_i(\bar{a}_m)$ . The structure of  $\underline{s}_i$  is induced by the same operation,  $\oplus$ , which relates  $\underline{p}_i$  to  $\underline{r}_i$  and  $\underline{s}_i$ .

This mode of sequence interpretation does *not* imply that the gradient phonetic output is necessarily a neatly temporally segmented sequence of the phonetic information associated with each segment. We know this is not true: acoustic cues and gestures overlap across segments in a way that makes it clear that segments are units that are, at best, strictly internal to the cognitive system, and are not “physically real.” It does not imply this for two reasons. First of all, the phonetic realization of a segment is crucially context-dependent in this model: a segment gets to affect the phonetics many times, once as  $z_i$  and then again for each  $z_{j \neq i}$  where it appears as part of the context. Secondly, although in the Chapter 3 models we translated each category in the surface representation into a single element of a phonetic representation sequence, more realistically we want to translate each category into a sequence of phonetic representation elements. Allowing for the output to have temporal overlap (“frame overlap” in the automatic speech recognition literature) would then have no additional architectural consequences whatsoever if the overlapping parts of two adjacent frames straddling an abstract segment “boundary” (meaning that the

first is considered part of segment  $i$  and the second is considered part of segment  $i + 1$ , even though they actually overlap in time) were always along independent dimensions of the gradient representation. In general, there will be consequences, as we will be required to say how the phonetic representations in the overlapping regions combine, (see Browman & Goldstein 1993 for a sketch of how this might work in production). Combination under overlap then also makes the inverse problem more complex. For the moment, however, the point is simply that saying that the “sequence” of segments is abstract with respect to the actual phonetic representation’s temporal ordering.

Finally, I should point out that the architecture itself does not say that  $\bar{x}_i$  needs to be a function of *only* the categorical AC-representation, but I will assume that it is. The alternative one might envision is that it also contains some information which is gradient (peripheral, phonetic—as in F0 transform in the sociolinguistic model presented at the end of Chapter 3). The consequence of this would be to set up potential conflicts between the output of one segment’s interpretation and the contextual input to another. We would be required to say something about how the different mappings for the different segments combine to resolve this conflict—most naturally, just compose them, but this would imply that the the translation to and from gradient phonetics takes place as if in multiple “stages,” and we would then need to say how those stages combine (that is, state an ordering). In what follows, I will assume that this is not the case. The reason for this is not empirical but rather a research strategy. The consequences of making this particular assumption are rich but simple to work out; the alternative theory subsumes this one, and, in the absence of any particular empirical motivation one way or the other, I hew to the theory that is logically prior. Although this might seem somewhat limiting, this formulation is clearly

sufficient to cover a wide range of different contextual phonetic effects. (In the second half of the chapter, we will see that some extensions of the architecture would force us to say that we can sometimes translate *back* from some gradient interpretation object(s) to their categorical representation, which would only be legitimate sometimes, not always; but this is different. See below.)

The crucial point about the AC-representation that makes it different from an SC-representation is that **the results of applying phonetic transforms are not coded in the AC-representation**. Rather, a segment which may have contextually-variant pronunciations is represented, alongside the contextual information, and phonetic interpretation combines these to give a single phonetic realization. An immediate consequence of this is that no contextual variant pronunciation that is the result of applying contextual phonetic transforms is be coded lexically as its own segmental category. In contrast, an SC-representation will code a different category for each contextual pronunciation variant of a segment. (We will see below that this actually forces the general shape of the architecture upon us.)

An analogy will help: suppose I am tuning a ukulele and I am given a slip of paper that tells me to tune the four strings to G4, C4, E4, A4. I adjust the tuning pegs to give a particular tension in the strings. Once I do this, four particular frequencies (392 Hz, 262 Hz, 330 Hz, and 440 Hz) will be encoded in the tension in the four strings. If I pluck the G string, a note with a fundamental frequency of 392 Hz will sound; conversely, if I play a loud 392 Hz tone in the room, the G string will be maximally resonant, and I can thereby identify the note that was played. Now, in a sense, the categories G, C, E, A are encoded in two places—in the strings of the ukulele, and on my piece of paper. This fact



remains true if I believe that ukulele tunings are subject to contextual variability—for example, it remains true even if my piece of paper also tells me how to adjust each string to the so-called “Canadian” tuning—A3, D4, F#4, B4: multiply the frequency of each by an equal-tempered whole tone,  $\sqrt[6]{2}$ , and additionally divide the frequency of the first string by an octave, that is, 2—should the environment demand it. There are still only finitely many tunings licensed by my piece of paper, and, even if a piece of paper of this kind went on forever, it could only ever give me a discrete subset of the physically possible ukulele tunings. However, there is nothing on the paper that says A3, D4, F#4, B4; rather, it gives me two-step instructions for obtaining this tuning, via G4, C4, E4, A4. Were I to play a loud tone in the room in an attempt to work out the single “paper object” corresponding to a particular string and its tension, I would turn up nothing. Of course, the two representations of the frequencies are also different types of things: if I want make another tuning adjustment, I need paper instructions, and if that tuning adjustment is dependent on the current state of the ukulele, then I need to first work backwards from the tension in the strings to obtain the tuning frequencies. If the only way that I can be given instructions for tuning is in terms of paper-type categories, then I will be stuck if I come up empty, which will happen when I am in the Canadian tuning, unless I know how to go all the way back to the “underlying” categories G4, C4, E4, A4. *Mutatis mutandis*, this means that there can be no feeding relations from strict allophonic outputs to categorical phonology, or from strict allophonic outputs to the categorical environments for further allophonic processes. When phonetic transforms apply, the resulting categories are epiphenomenal with respect to the categorical phonology. See below and Chapter 5 for the implications of this.

Given that the individual contextual variants of a segment have no categorical repre-

sentation, we can see that applying phonetic transforms will naturally give rise to complementary distribution: any alignment of the output of transforming category *A* with some realization of a separate category *B* is accidental, and should result in pronunciations which only ever occur in a particular restricted environment (this will be discussed further in the next section and in Chapter 5). In contrast, a string transformation from *A* to *B* should have a result that behaves exactly like the category *B*, even if *B* is subject to idiosyncratic phonetic adjustments for each environment in which it appears: if the realization of a particular [t] is simultaneously subject to the phonological instruction “change to [d] due to a following voiced segment” and the phonetic instruction “front by a particular amount due to a preceding [i],” the degree of fronting should be that of a [d], not of a [t], if the two are different, and the base phonetic representation to which this fronting applies should be exactly that of a [d] in another environment. A particular phonetically-adjusted realization of [t] should appear in one particular environment and nowhere else, because it cannot be coded directly in the lexicon and can only be obtained by applying phonetic interpretation, but a [d] derived from a [t] (fronted or not) will not exclusively occur before voiced segments, but will be the same as lexical or otherwise-derived [d] in other positions. To reiterate: a phonetically-adjusted realization of [t] “cannot be coded directly in the lexicon” not because there is an explicit grammatical constraint on what categories can be coded in the lexicon, but because it is not the right type of object; this results in complementary distribution. See Chapter 5 for expanded discussion of the issues under interactions of phonetic processes.

We have therefore now answered the question of how phonetic transforms relate to, and account for, complementary distribution: the phonetic (oracle-) categories that are

obtained are not coded categorically, and cannot be reproduced for another segment or in a contradictory environment except by accident. Conversely, oracle-categories which are not in complementary distribution cannot be related by phonetic transform. This account of complementary distribution aligns with the classical notion of “strict allophony” in the sense that the contrasts between the various phonetically transformed realizations of a category are not coded lexically, and are thus “non-contrastive.”

Whether phonetic transforms in fact *will* account for any particular cases of complementary distribution traditionally falling under the umbrella of “allophony” is another story; however, suffice it to say that they can, and, to preview Chapter 5, they should, all things being equal, provided there is enough of the right kind of evidence available to the learner.

#### 4.1.3 Problematic and unproblematic appeals to surface representations in phonological theory

The absence of SC-representations under this theory is in conflict with the premises of certain arguments and proposals in phonology. In this section I document some highlights of which appeals to surface representations conflict and which do not; the “surface representations” that do not conflict could just as easily be AC-representations, and are therefore perfectly acceptable under the current architecture.

#### 4.1.3.1 No problems with AC-representations

Let us begin with a straightforward example of the reasoning; this case will turn out *not* to be a problem. McCarthy 1986 presented some convincing evidence for the *Obligatory Contour Principle* (ultimately rooted in a proposal of Leben 1973). The principle unifies a class of exceptions to rules and restrictions on possible lexical items across languages by disallowing certain sequences of identical segments. Some cases are given in (97) and (98):

(97) *Afar* (Bliese 1981)

$$\text{Syncope Rule: } \left[ \begin{array}{c} \text{V} \\ \text{--long} \\ \text{--stress} \end{array} \right] \rightarrow \emptyset / \#(C) \left[ \begin{array}{c} \text{V} \\ \text{--long} \end{array} \right] C_{\alpha} - C_{\beta} \text{V}$$

except if  $C_{\alpha} = C_{\beta}$

*default case:*

$\text{ħam}\underline{\text{il}}\text{a} + \emptyset$  (*swamp grass +acc.*)  $\rightarrow$   $\text{ħam}\underline{\text{l}} + \text{i}$  (*swamp grass +nom.*)

$\text{di}\underline{\text{j}}\text{b} + \text{te}$  (*marry + 3.sg.fem*)  $\rightarrow$   $\text{di}\underline{\text{b}} + \text{e}$  (*marry + 1.sg*)

*exceptional case:*

$\text{mid}\underline{\text{.ad.}}\text{u} + \emptyset$  (*fruit+acc.*)  $\rightarrow$   $\text{mid}\underline{\text{.ad.}} + \text{i}$  (*fruit +nom.*)

$\text{al}\underline{\text{al}} + \text{te}$  (*race +3.sg.fem*)  $\rightarrow$   $\text{al}\underline{\text{al}} + \text{e}$  (*race +1.sg*)

(98)

*Tonkawa* (Hoijer 1949, Kisseberth 1970)

Syncope Rule:  $V \rightarrow \emptyset / \#CVC_{\alpha} - C_{\beta}V$

within stems, except if  $C_{\alpha} = C_{\beta}$

*default case:*

notoxo- (*hoe, stem*)  $\rightarrow$  notx + o (*hoe+3.sg*)

pitena- (*cut, stem*)  $\rightarrow$  pitn + o (*cut+3.sg*)

*exceptional case:*

hewawa- (*die, stem*)  $\rightarrow$  hewaw + o (*die+3.sg*)

hamama- (*burn, stem*)  $\rightarrow$  hamam + o (*burn+3.sg*)

The Afar syncope rule deletes the second of two short vowels in open syllables at the beginning of a word, if it is unstressed and not word-final. (An unrelated process deletes stem-final short vowels before the nominative *i*-suffix.) An exception is made when this vowel falls between two identical consonants, however. Similarly, in Tonkawa, a syncope rule deletes the vowel in the second of two CV syllables when it is not word-final (an arguably separate process deletes the first of a sequence of two vowels). Again, an exception is made when the vowel falls between two consonants. McCarthy links this to a restriction on Arabic triconsonantal roots: only the second two consonants, but not the first two, can be identical (there can be stems like *samam*, “poison,” but no stem like \**sasam*). The explanation assumes autosegmental theory: rather than being a sequence of segmental feature bundles, (see Chapter 3), a lexical item is actually a graph, where

features are “associated” with each other, and, as a special case, associated with “timing units” that give rise to the segmental ordering. Importantly, a single set of consonant features can be associated with two different timing units, in this case, to two consonants. The Obligatory Contour Principle bans having two adjacent segments with the same features unless it is because they are sharing the same feature bundle, where “adjacent” in Arabic means adjacent consonant-wise. Sharing of features is highly restricted, and in Arabic is limited to the second two, not the first two consonants. While this might seem baroque, McCarthy presents evidence from language games that in *samam*-type stems, the second two consonants behave as a unit (and the interpretation of “adjacency” with respect to just consonants is needed independently throughout Semitic morphology). In Afar and Tonkawa, adjacency is interpreted segment-wise, and the rules do not have the power to “restructure” the associations between timing units and features after they apply; thus syncope would give an output violating the OCP, thus it fails to apply. Thus the explicit condition on both rules can be dropped, assuming that the behavior of the rule when its output would violate the OCP (skip the rule) is well-defined.

The OCP is a classic case of an *output filter*. What is crucial here is not that it the OCP is a filter, but rather that it makes reference to outputs. Rather than putting a condition on the input to the rule, a condition is placed on the output that determines whether the rule will apply. This is not forced by the Afar or Tonkawa patterns, but it is necessary in order to unify the condition with the Arabic triconsonantal root restriction under the OCP, because, given the segment-wise interpretation of “adjacency” in Afar and Tonkawa, the only sense in which there are adjacent identical segments banned in either is in the output. This is not merely a restriction on surface representations: it is

a restriction on *all* representations—but it therefore extends to surface representations. If the surface representation is an SC-representation—a surface representation making use of non-contrastive features, to be reinterpreted here as phonetic—it raises a potential problem under the current theory.

Since the crucial question is whether the OCP can be evaluated, the question is whether the OCP can be verified after the rule applies. In the original formulation of the Afar and Tonkawa rule exceptions, restrictions on the forms that the rule applies *to*, this is not an issue; the features are present in the AC-representation, which is the input to phonetic interpretation, and thus it would be fine if these rules were phonetic no matter what. In this case, there is little reason to worry: the deletion of a segment is in no way a good candidate for reinterpretation as “non-contrastive feature change.” Since there is no independent reason to think that either rule is phonetic,<sup>1</sup> I will assume that there is no problem here; but this is the kind of problem we would be looking for.


It comes up under other analyses as well. Gouskova 2003 reanalyses Tonkawa syncope under Optimality Theory. In this analysis, the environment for vowel deletion is deduced as being on those vowels which would permit an ideal trochaic footing of the string (as indicated by Phelps 1975, the syncope is more general than the second syllable, and the more general rule reapplies from left to right). This does two things: first, it gives

---

<sup>1</sup>With one exception: according to Bliese, in the Northern Afar dialect, there are certain lexical exceptions to the syncope rule, such as the verb *alif*, “to close”: for the first person singular, we get *alif+e* rather than *alf+e*. As we will discuss in the next section, we do not expect to find lexical exceptions among phonetic rules. However, apparently, the vowels that escape deletion in this way are often reduced. Although Bliese only presents this as a passing remark, this reduction could be attributable to a phonetic rule, if it were found to be phonetically shorter or more centralized than the same short vowel in a non-syncope (but otherwise maximally similar) position. In this case, there should be no OCP effects, but such a prediction is moot anyway, as reduction (as opposed to deletion) should not be sufficient to trigger even a phonetic OCP effect. Besides this, the language would need to be such that these exceptions happened to coincide with cases where the OCP would be relevant.

an account of the environment for the rule in terms of constraints falling out of a metrical analysis; second, this environment is derived by way of output constraints: the string is adjusted to accommodate an ideal metrical analysis. As discussed in Chapter 1, this is in fact the only grammatical mechanism standardly made available in Optimality Theory.

The consequence of this is that the Optimality Theory analysis both changes the representation of the grammar, (relevant primarily to learning, as discussed above), and fails to assert the sequence-of-ordered-representations structure that can be induced by the composition of operations: the grammar does not say how the actual computation of the candidate output forms proceeds. However, what is crucial here is that this analysis induces a still stronger appeal to output representations, because, however the computation proceeds, the output must include both the result of deletion and the relevant metrical analysis. Take the evaluation for [notxo], “he hoes,” for example:

/notoxo + o/		RhType=T	Stress-to-Weight	Max(V)	*Clash
(99)	 (nót)(xó)			**	*
	(nó.to)(xó)		*!	*	
	(no.tó)(xó)	*!	*	*	*

In the correct output candidate, a vowel is deleted (in fact, two); but, in order to evaluate this candidate, (for example, to determine where and whether the RhType=Trochaic constraint is violated by having moras grouped together into feet with stress on the second one, and where and whether the Stress-to-Weight principle is violated by having a light syllable stressed), we need to be able to examine the stresses and feet. An operation of deletion feeds an evaluation that makes reference to metrical structure. The analysis then evaluates the output of syncope to determine where and whether syncope should apply.



If we were to find reason to think that syncope should be treated as phonetic under the current architecture, it would be extremely problematic. Feet are entirely abstract and absent from the phonetic representation, and, although stress and syllable weight do have gradient phonetic correlates, the RhType=Trochaic and Stress-to-Weight principles assign violation marks on the basis of categorical differences (heavy/light, stressed/unstressed). The output representation being evaluated would thus need to be both gradient and categorical, which, according to this architecture, is impossible. This contrasts sharply with the analysis of Tonkawa given by Noske 1993, which treats syncope as the result of removing unsyllabifiable vowels. Although the analysis makes heavy use of output constraints to derive the optimal syllabification, this syllabification can be (in fact, crucially is) fully determined on the basis of the non-syncopated segmental representation; unlike in the OT analysis, no syllabification or other suprasegmental analysis of syncopated outputs needs to be done. Thus no problem would be posed if syncope were phonetic.

To sum up: grammatical output constraints can in principle be in conflict with the absence of the SC-representation—but not all analyses making use of a categorical surface representation are problematic: **(i), if there is no reason to attribute any of the relevant processes to the phonetic component, an AC-representation is sufficient**; and, **(ii), to be “relevant,” a process needs to feed the evaluation of the output constraint, or, more generally, to feed further (categorical) grammatical computations**; otherwise there is no potential conflict.

#### 4.1.3.2 Historical changes

We now turn to cases where one or both of these issues clearly does arise. Padgett 2003 discusses *post-velar fronting* in Old East Slavic and its synchronic residue. Modern Russian is generally understood to have a lexical contrast between palatalized and plain (arguably velarized) consonants.

	Labial		Dental		Postalveolar	Palatal	Velar
Stop	p	p <sup>j</sup>	t	t <sup>j</sup>			k k <sup>j</sup>
	b	b <sup>j</sup>	d	d <sup>j</sup>			j
Fricative	f	f <sup>j</sup>	s	s <sup>j</sup>	j:		x x <sup>j</sup>
	v	v <sup>j</sup>	z	z <sup>j</sup>			
Affricate			ts		t <sup>j</sup>		
Nasal	m	m <sup>j</sup>	n	n <sup>j</sup>			
Lateral			l	l <sup>j</sup>			
Rhotic			r	r <sup>j</sup>			
Glide						j	

Table 4.1: The consonant and vowel inventory of Russian.

Early Common Slavic had no palatalization at all. A change took place that took  $k, x > t, s$  before long and short *i* and *e*, (which were the only front vowels), as well as *j*, and then the appearance of velars became perfectly predictive of the following vowel being back. The Old East Slavic post-velar fronting was a diachronic process by which the vowel [ɤ] fronted to [i] following velars. Padgett suggests that this was because there was no contrast between [ɤ] and [i] following velars, and, under the pressure to maximize phonetic contrasts posited under *Dispersion Theory*, (Flemming 1995), [ɤ] fronted in order to be as far away from the high back vowel [u] as possible. This was not possible following non-velar consonants, say [p], according to Padgett's analysis, because there it was possible for [i] to appear, and a dispreference for merging lexical distinctions made such a fronting


undesirable. (Confusingly for us, the first change was called the *first palatalization*; this is confusing because this was *not* the change that resulted in the palatalization contrast that is evident from the table, and that we are going to need to talk about below. This happened much later, but before the post-velar fronting. The first palatalization was instead a change that changed velars to palatal fricatives, historically by way of a pronunciation with palatalized secondary articulation.)

Padgett’s analysis is done using Optimality Theory. Deviating from standard assumptions, Padgett assumes that the OT computations operate on entire languages, not just to compute the pronunciations of individual morphemes or utterances. What this means cognitively—what cognitive process this is supposed to correspond to—is not entirely clear. Each one of Padgett’s tableaux is evaluated over a small collection of multiple idealized “words,”  $pi, p, k^i$ , and so on, with each candidate listing an output for each one of the words. He says that such a collection of words is an “idealized language.” The reason he must include multiple items is clear: Dispersion Theory requires a comparison among different forms, (for example, it is the presence of  $[u]$  in some forms in the language that forces  $[\text{ }]$  forward, not just facts about  $[\text{ }]$ , nor any lexical representation in which it appears, nor any surface pronunciation). Flemming 1995, in introducing Dispersion Theory, does something similar, in using OT computations to evaluate “systems of contrasts,” rather than individual underlying–surface pairs. In Padgett’s analysis, this is made necessary because of the presence of a family of constraints called  $\text{Space}(\text{color}) \geq \frac{1}{n}$ , each of which assigns a violation for every pair of items in the (complex) output candidate that differs only in a vowel contrast spanning less than  $\frac{1}{n}$  of the “color” dimension for vowels, corresponding roughly to the second formant. For example, if  $pi, p$ , and  $pu$  were all in


the candidate, then  $\text{Space}(\text{color}) \geq \frac{1}{2}$  would assign two violations,  $\text{Space}(\text{color}) \geq \frac{1}{1}$  would assign three, and  $\text{Space}(\text{color}) \geq \frac{1}{3}$  would assign none. For the candidate containing only  $\text{pi}$ ,  $\text{t}$ , and  $\text{pu}$ , only  $\text{Space}(\text{color}) \geq \frac{1}{1}$  would be violated, and it would assign one violation. These multi-form candidates are also necessary for the \*Merge constraint, which assigns a violation for each pair of non-distinct output candidate elements that correspond to distinct inputs: the input  $\text{pi}_1, \text{p}_2$  would incur a violation if paired with  $\text{pi}_1, \text{pi}_2$ , but not with  $\text{pi}_1, \text{p}_2$ , where the subscripts indicate which inputs correspond with which outputs. I will interpret Padgett's multiple-element candidates as implying that the computation of a surface form for any given input representation needs to simultaneously take into account the implications of the grammar for every other possible input representation. That is, I will interpret Padgett's tableaux as representing part of the normal phonological grammatical computation for a form, as opposed to some other, unspecified part of linguistic cognition; the only difference from the normal conception is that, to operate the phonological grammar one needs to operate on all forms at once. This is consistent with everything that Padgett says.

Here are the computations Padgett gives for the earlier and the later stages of Old East Slavic, before and after the post-velar fronting came into the language. The Space constraint is  $\text{Space}(\text{color}) \geq \frac{1}{2}$ .

(100) *Before post-velar fronting:*

$p^{i_1} p_2 p_{u_3}$ $k_5 k_{u_6}$ $\psi_{i_4}$	*Merge	Id(color)	Space
 $p^{i_1} p_2 p_{u_3}$ $k_5 k_{u_6}$ $\psi_{i_4}$			***
$p^{i_1} p_2 p_{u_3}$ $k^i_{i_5} k_{u_6}$ $\psi_{i_4}$		*!	**
$p^{i_1} p_2 p_{u_3}$ $k^i_{i_5} k_6$ $\psi_{i_4}$		*!*	***
$p^{i_1,2} p_{u_3}$ $k^i_{i_5} k_{u_6}$ $\psi_{i_4}$	*!	**	
$p^{i_1,2} p_{u_3}$ $k_5 k_{u_6}$ $\psi_{i_4}$	*!	*	*

*Post-velar fronting grammar:*

$p^{i_1} p_2 p_{u_3}$ $k_5 k_{u_6}$ $\psi_{i_4}$	*Merge	Space	Id(color)
$p^{i_1} p_2 p_{u_3}$ $k_5 k_{u_6}$ $\psi_{i_4}$		***!	
 $p^{i_1} p_2 p_{u_3}$ $k^i_{i_5} k_{u_6}$ $\psi_{i_4}$		**	*
$p^{i_1} p_2 p_{u_3}$ $k^i_{i_5} k_6$ $\psi_{i_4}$		***!	**
$p^{i_1,2} p_{u_3}$ $k^i_{i_5} k_{u_6}$ $\psi_{i_4}$	*!		**
$p^{i_1,2} p_{u_3}$ $k_5 k_{u_6}$ $\psi_{i_4}$	*!	*	*

The crucial difference in the two grammars in (100) is the ranking of the Space constraint. The ranking  $\text{Ident}(\text{color}) \gg \text{Space}$  puts faithfulness above the need for dispersion, thus barring post-velar fronting, while the reranking  $\text{Space} \gg \text{Ident}(\text{color})$  allows dispersion to take effect. The existence of a three-way contrast elsewhere blocks any movement of the central vowel forward, as this would violate \*Merge, but, after velars, there is no contrast, which frees the reranked grammar to front  $[k]$  to  $[k^i]$ .

There are several things which are sources of potential conflict between this analysis and the architecture proposed here. As before, the two sticking points for any categories that are proposed to be in the surface representation is whether they are in complementary distribution and whether they (fail to) correspond to any lexically-specified categories. As always, I leave the worked-out explanation of why these are problems for Chapter 5, where it will be shown that the grammatical treatment of patterns satisfying either of these two notions of “allophone” as phonetic is not actually a hard restriction, but rather a

preference; I will specify what the factors that go into the assessment are, given a rational learner. For the moment, however, I continue to simply assess what would go wrong when these patterns are indeed treated as phonetic.

Here, we may note two things like this. First, the distribution of [j] and [i] is complementary. Although it is not shown, a high-ranked constraint, Palatalization, enforces palatalization of all consonants before front vowels. At this stage of the language, the constraint actually shows its effects across all consonants, not just velars, a point which I will discuss in a moment. This means that the appearance of [j] or [i] is actually totally predictable, because [i] will always be preceded by a palatalized consonant. Second, the distribution of [k] and [kʲ] is also complementary, because [kʲ] only appears before front vowels, and [k] never appears before front vowels, and more generally this is true across the velars. This is actually not true for the other places of articulation, despite the fact that I show what appears to be a consistent complementarity for the non-velars, represented by [p], in (100) in the input forms (I deviate slightly from Padgett here in indicating the palatalization on the non-velars—see below). This is because of the deletion of the high short front vowel in certain positions, (the “front yer,” written Ъ in historical Slavic linguistics and probably pronounced [ɪ]), which left some instances of palatalized non-velars followed by something other than front vowels; the first palatalization had already protected the palatalized velars from this fate, by lexically eliminating the possibility of velars before [j] (and eliminating velar–front vowel sequences across the board). This too will be discussed shortly.

In assessing the question of lexical contrast—that is, whether it is fair to say that [k]/[kʲ] is actually not marked lexically, thus providing another reason to judge the differ-

ence to be a purely phonetic one, not appearing in the AC-representation—we must be careful. This is because Padgett appeals to the principle of Richness of the Base, which is just a name for the conventional architectural assumption in Optimality Theory stating that the grammar serves only to constrain the mapping between lexicon and phonetics, not to place restrictions on what is allowed to be in the lexicon (for example, statements of lexical inventories). The nearest equivalent to asking whether a category can be marked lexically is ordinarily to ask whether it actually is marked lexically. However, this does not work here, as the interpretation of Richness of the Base is complicated by Padgett’s appeal to complex inputs that contain all lexical items. Richness of the Base would imply that the evaluation should always be over an input–output pair that contains not every possible lexical item, not just every actual lexical item (whatever this means in Padgett’s idealized language). Since this would make the Dispersion Theory analysis Padgett gives impossible to maintain,<sup>2</sup> he states that all of his analysis takes place in the post-lexical component, and the restrictions on inputs are due to neutralizations in the lexical phonology (see below in this chapter for more discussion of this distinction and its status in the present theory). The relevant question for Padgett is therefore whether a distinction is present at

---

<sup>2</sup>In the case of (100), this means that *kʲi* (among infinitely many other things) also needs to appear in the input, and thus have a corresponding item in the output. This will incur a violation of \*Merge for the candidate that is currently shown as winning in the second tableau, with post-velar fronting; but then the output candidate on the last line, with fronting of [p] to [pʲi], will be no worse by \*Merge. Space will be no better off, because the addition of *kʲi* will mean that two violations are incurred among the velar-initial elements, too, and it will not be to the advantage of the winning candidate that there is an extra “space” to fill left by the absence of *kʲi*: richness of the base would imply that there is no such space. One response might go like this: “A higher-order Richness of the Base principle is at work. In this theory, the grammar needs to assign a consistent output across all possible subsets of the lexicon, when each is submitted for evaluation. The selected input sets are just particular examples.” This *really* will not work, however: aside from the fact that adding the whole lexicon back in will give inconsistent results, as indicated in the text, subsets of the sets Padgett gives are also problematic: if we exclude [pʲi] from either of the input candidates Padgett gives, then the crucial violations of Merge that rule out fronting [p] are also alleviated, leading to another inconsistent result.

the output of the lexical phonology. This is also sufficient for us: if it has already been determined that the relevant grammatical changes operate on representations for which the [k]/[kʲ] distinction is not marked, then it would seem quite clear that this distinction could be interpreted as arising in the phonetic component. This line of reasoning will be worked out in further detail later in the chapter, and an analysis of the post-velar fronting phenomenon (albeit one which does not in fact look like this) will be given momentarily.

First, however, I address one side issue that comes up in Padgett's analysis: the Space constraints make crucial reference to the phonetic interpretation of particular features. It is not totally unreasonable to think that the information operated over in the categorical phonology might have some general phonetic content which could be evaluated in something like the way suggested here. As we discussed in Chapter 3, it is at least plausible to assume some general constraints on the interpretation of particular features, even if these are vague and subject to further specification in the phonetic component (and in fact, such constraints are the minimum needed in order to say that features are not simply classificatory); however, in the categorical phonology, Space amounts at any rate to simply penalizing particular combinations of categories. To give the "fraction of the total width" dimension real force, it would be reasonable to instead propose that the Space constraints are part of the phonetic component, which, again, undermines the possibility of doing the analysis over surface representations, and would lead us to question the status of the SC-representation. However, the constraints would need to be substantially reformulated: as stated, in the phonetic grammar, they should be satisfied by small phonetic adjustments, not only by the wholesale fronting of [ ] to [i].

I now discuss how this case would be handled if some or all of the relevant gram-



matical operations were taken to be phonetic. The [j]/[i] alternation is a good candidate for a phonetic transformation. The distribution of the two is complementary, and the distinction is not crucial to any lexical contrast, nor to triggering any phonological process (consider Padgett's assertion that the alternation can be handled in the post-lexical phonology). This becomes still clearer when we consider the fact that the complementary distribution of [j]/[i] holds across all the other consonants, too, not only the velars. This is still true today in Russian, and the traditional analysis of this fact in Russian is that the palatalization contrast, which is phonemic, (albeit somewhat marginal for velars where it is mainly contrastive in loanwords), and triggers an allophonic change from [i] to [j] in the presence of a plain consonant. There must be some such process—that is, there must be cases where an underlying [i] surfaces as [j]—given morphological alternations like *pot+i* → *pot*, *sweat + nom/acc pl.* (versus *putʲ+i* → *putʲi*, *road + nom. pl.*). The analysis of Dresher 2009b, following that of Jakobson 1929, is that the distinction became non-contrastive following another change in Old East Slavic, the shift from non-contrastive to contrastive palatalization. (This was alluded to above, and will be explained in a moment.) This is slightly different from Padgett's analysis in asserting that the [j]/[i] distinction stopped being coded lexically, although both capture the distributional pattern. This is what we will assume here. Dresher further supposes that the underlying representation of this vowel was then as [−back], and that the alternation in the vowel was then attributed by learners to exactly the rule just described, with plain consonants, marked as [+back], spreading this feature to following [i] to yield [j].

This does not account for the post-velar fronting, since the language had only [j] following velars, and evidently had only plain, and not palatalized, velars in this position.

Thus the presence of plain velars would seem to have supported the maintenance of these plain-[] sequences, given this rule, not a shift to [kʲi]. Where Padgett explained this shift by a pressure for vowel dispersion, however, tolerated only in the velars due to the lack of a palatalization contrast, Dresher explains the shift using a version of Contrastive Specification (see Chapter 3). The absence of a palatalization contrast in the velars, following the first palatalization, led to their becoming underspecified for the [back] feature. This implied that they could not trigger the backing that the other plain consonants triggered. Thus the vowel surfaced in its underlying [−back] form.

This would be one possible analysis—simply translate the rule backing [i] to a phonetic rule triggered by a [+back] feature—if we assumed a Contrastive specification account. We would then still need to account for the palatalization of velars in this position. Padgett accounts for this by highly ranking a Palatalization constraint (asserting, in Dresher’s featural analysis, that if consonants are followed by [−back] vowels, they must be [−back]); it is ranked over both Ident(pal) (i.e., Ident(back)) and \*kʲ. In fact, this is the same ranking that Padgett uses to account for a precursor to the first palatalization in an earlier stage of the language, an ranking which, as he points in a footnote, was also extended to \*pʲ, (where p stands in for all the non-velars), by the time of the post-velar fronting. This secondary palatalization on all consonants is in fact very important, as Dresher points out, because it was this palatalization that is understood to have been reanalyzed as phonemic just prior to post-velar fronting: [], which also triggered the alternation, was substantially reduced in certain positions, and eventually disappeared (dn → dʲnʲ → dʲnʲ, *day*: word-final position is a reduction position and the immediately preceding syllable is not). Some residue of this palatalization process still exists today, triggered by [e], (sʲirot+e → sʲiratʲe,

*orphan* + *dat.*, versus *sʲirot*+*i* → *sʲirat*, *orphan* + *gen.*), plus a handful of idiosyncratic suffixes. Dresher assumes that the more general palatalization persisted, but, at the stage of post-velar fronting (and since) was crucially ordered after the rule backing [i]. That rule has no effect following velars, because they are underspecified for [back], but for other consonants bleeds palatalization.

However, neither of these explanations deals directly with the fact that the grammar inducing the [kʲi] sequences would not be assessed as a good fit to the phonetic evidence available to the learner at the time of the post-velar fronting, which evidently has all those sequences as [k]. A better fit to the data could be obtained, under Padgett's analysis, by ranking \*kʲ high, and, under Dresher's analysis, by filling in the missing [+back] feature prior to the backing rule, or analysing backness as contrastive for velars, with the presence of [k] being a more decisive factor in coming to this conclusion than the absence of [kʲ]. Why should any of these legitimate analyses have been dispreferred by learners?

I conclude that they simply would not have been, and the alternative opened up by the phonetic approach is a way of saying that the shift was initially *gradient* and thus potentially *gradual* in time. This is important, because the emphasis throughout this dissertation has been on the *tradeoff* in learning between goodness-of-fit and goodness-of-grammar (e.g., simplicity considerations). In order to justify deviations from fit to the data, the grammars that fit the data must be correspondingly dispreferred (at least—or they might even be impossible). Since the data that the learner is attempting to explain at this stage consists of [pʲi], [p], and [k], but crucially no [kʲi], the analysis should not deviate too much from fitting the data unless there is a corresponding increase in the inherent preferability of the grammatical analysis proposed.

An analysis which is both gradient and gradual goes as follows: the learner first arrives at an analysis which does not deviate from this pattern at all, and thus, at this stage, the language is rather different from what has been previously proposed. There is a phonetic  $[\text{ }]/[i]$  rule, either fronting, backing, or deriving both from an archiphoneme. Backing provides a better fit to the data, because (for independent reasons) there is no word initial  $[\text{ }]$ , but there is word-initial  $[i]$ , and so, presumably, any single-feature environment for fronting will compromise the distribution somewhat. Whatever residue of palatalization can be maintained for non-velars, on the other hand, must be a categorical rule, and thus must feed the  $[\text{ }]/[i]$  rule. Crucially, therefore, this alternation must exclude the underlying representation of  $[\text{ }]/[i]$  ( $/i/$ ) as an environment, or else the distribution that forms the basis for the  $[\text{ }]/[i]$  alternation would be undone. At this stage, then,  $[k]$  is maintained.

Once we have specified the velars as active triggers for the allophony, we can adapt the contrast-based ideas of the previous analyses, which both rely on the absence of a palatalization contrast for velars to tolerate the innovative  $[i]$  pronunciation in that environment, to motivate a shift in the appropriate environment. However, rather than attempt to use the relative goodness of grammars to do the work, I propose that functional pressures were at work. In particular, the goal of fronting was *phonetic enhancement* in the direction of the underlying  $/i/$  (Keyser & Stevens 2006). This implies that speakers gradually both backed and fronted  $/i/$  after velars.

Before filling in the detail suggesting why this should lead to an eventual reanalysis as  $[k^j i]$ , rather than  $[k]$  followed by backed and fronted  $[i]$ , I point out that, while there is no good theory of what conditions should license or encourage a particular phonetic enhancement of a segment, we can put forward two suggestions: (i) the unenhanced pro-

nunciation should be perceptible to the speaker as a non-prototypical realization of some AC-category; (ii) the enhanced pronunciation should be a more prototypical realization of that same category. The first condition is complex; it says that the unenhanced pronunciation should be perceived as a realization of some particular category, but it should also be perceived as deviant. This will not be the case for most processes. It will be the case for phonetic processes in domains where gradient acoustic perception is reasonably good, such as vowels or fricatives; however, since allophony, for whatever reason, seems to reduce perceptibility when it occurs, (see Chapter 3), more than just allophony will likely be required to be perceived as non-prototypical; the auditory apparatus must be independently equipped to detect the relevant acoustic differences, and the best case of this will be, as in this case, where there is an existing contrast: differences on the high end of the F2 dimension, which cue both palatalization and the backing of [i], (Hamann 2003), are needed independently to detect the contrastive palatalization difference in the non-velars. (It should also be the case that sociolinguistically stigmatized neutralizations will be sensed to be non-prototypical, albeit in a very different sense. However, this is the nature of functional pressure: it is a confluence of disparate forces pushing the grammar in a particular direction.) The second condition says that the enhanced pronunciation should be more prototypical; for this to take effect, all the same conditions hold as for the first condition. However, notice that this also does the work of the \*Merge constraint: to the degree that the percept is also perceptible as a good realization of some *other* category, or sequence of categories, then, through the joint probability, the speaker's degree of belief in a prototypical representation *of the target category* is also decreased, no matter how prototypical it may actually be.

Why should speakers make gradient changes to satisfy these pressures, rather than categorical changes? Furthermore, isn't the idea that speakers are making changes to F2 ambiguous as to whether it is the vowel or the degree of velar palatalization changing, given what we just said? Speakers, of course, cannot make changes in F2 directly, and, under reasonable assumptions about production, will be able to manipulate palatalization and vowel frontness somewhat independently; if so, under these assumptions, speakers must enhance vowel quality and not palatalization: there is no reason to think that the velar tokens would be perceived as anything but prototypical in this context. The changes will be gradient rather than categorical, however, because, given a large change in fronting, the velar *would* stop sounding prototypical, due to the overlap in acoustic cues.

In sum, unlike in many cases of allophony, speakers should have been sensitive to the phonetic difference. They would therefore attempt to enhance to get just enough fronting so that allophonic [k] sounds like [ki] without sounding too much like [k<sup>i</sup>i] (notice that it does not matter that palatalization is not contrastive for velars, just that it is detectable). Despite the phonetic backing process, speakers innovated a process of phonetic fronting applying to the same vowels, but only after velars; for learners, the two processes combined by  $\oplus$ , but eventually the cue became highly ambiguous, and a grammar generating [k<sup>i</sup>i] became preferred (see Chapter 5). A new categorical palatalization rule was introduced, palatalizing velars before [i], thus lifting the need to back.

While this explanation admittedly suffers from the lack of a full model of speaker innovation, we do have a reasonably well-developed learning model; given the right likelihood functions to learn palatalization and consonant categories, and the necessary improvements to the variable selection sampling scheme, the category and allophony model

discussed in Chapter 3 will be usable in testing these predictions largely unchanged. However, much more work is needed in incorporating categorical processes into the model. The trading relations I discuss here and in Chapter 5 depend heavily on what precisely the prior for phonological grammars looks like.

To sum up, we have now seen a case where the assumption of an SC-representation was crucial to the analysis, that of Old East Slavic post-velar fronting. However, in addition to being incompatible with our current assumptions, I took this analysis to be unsatisfactory, in that it proposed a sudden move to a highly innovative grammar, a problem shared even by previous rule-based analyses which would just as easily have tolerated AC-representations. I thus proposed a new account, suggesting a gradual drift towards the innovative system by means of additive phonetic transformations. Although this suggestion is still somewhat informal due to the lack of a full learning model, (an issue which is nevertheless shared by all other accounts), it is much better specified under the current assumptions than it might be otherwise, because we have a precise notion here of what it would mean to have a phonetic rule doing the enhancement.

#### 4.1.3.3 Opaque allophony

I finish this section by simply pointing out that it is not only unorthodox global-evaluation theories like Padgett's analysis that demand SC-representations. Any account of a strict allophony-type pattern done using categorical phonological grammar is equally a candidate for reanalysis that would force AC-representations, although many would be substantially less interesting than the Old East Slavic case. I list a few of these presently.

However, it is important to highlight one special case, namely, analyses of opaque patterns which attempt to reanalyse the pattern to eliminate the opacity (we called this “obscured complementary distribution” above; we are about to go through exactly that case, shortening/centralization/raising—henceforth “raising”—and flapping).

An allophonic process in principle either (i) can be obscured in its environment or in its output by the action of some other process or (ii) can obscure by its action the environment or output of another, previously applied, process. We will review the implications of the current architecture in Chapter 5, but, suffice it to say, the prospects are better than under monostratal Optimality Theory, which cannot get the composed relation to arise out of a combination of the two relations in cases like this. To make this less abstract, consider the case of raising and flapping. In that example, we saw that some English speakers pronounce the diphthongs in *write* and *ride* as [rjt] and [rajd] respectively. These diphthongs have an allophonic analysis based on the voicing of the following segment, but the complementary distribution is obscured. This is because there is another allophonic process, flapping, which makes *writer* and *rider* come out with (what appear to be) identical consonants following, [rjər] and [rajər] respectively. To get this kind of interaction between two processes arising from independent constraints is impossible in standard monostratal Optimality Theory (for an explanation, see McCarthy 1999). Mielke, Armstrong & Hume 2003 begin with the possibility of not treating the obscured raising distribution as being due to allophony, but rather just due to a static phonotactic fact. Recall from above that this is a standard tack taken on the Lexicon Optimization hypothesis. In many cases, this almost works, as most of the supposed alternations are actually morpheme-internal, but the grammar still needs to handle what alternations there are across morpheme boundaries.



In this case, that would be a problem, at least if the patterns were to be captured in the natural way: a monostratal grammar built out of surface constraints cannot be built by enforcing the surface generalization that [j] appears before voiceless consonants and [aj] occurs before voiced consonants, because it is not always true: the flap is voiced. In this case, however, the claim is that there are no such alternations, because the diphthong-flap boundary is never at a morpheme boundary (notice that the [t] in *writer*, flapped or not, is just as morpheme-internal as the [t] in *write*). Thus the opaque grammar simply never arises. Putting aside whether this is correct, (Idsardi 2006 presents evidence that it is not true in Canadian English: *i-th, y-th* → *jθ, wjθ*, not → *ajθ, wajθ*), the problem is still a serious one in general.<sup>3</sup>

This does not turn only on the question of AC-representations versus SC-representations: the nature of OT grammar itself is also potentially at issue, and this kind of problem arises whenever there are two processes interacting in particular ways, allophonic or not. However, the issue of how to specify the grammar is still directly tied to the representation of allophony. If the obscuring process (in this case, flapping) is treated in a second mod-

---

<sup>3</sup>Another possibility which is always available in principle is to deny that the opaque case emerges in the same way as the transparent cases. Some effects can be derived in this way in Optimality Theory using *local constraint conjunction*: in particular, cases of counterfeeding order, where a rule does not apply in what appears to be the correct configuration because that configuration was actually generated by a later rule. This leads to underapplication with respect to the surface environments, and so additional markedness constraints can simply be added to the grammar to suit the particular nuances of the environment in which the rule fails to apply as expected. The solution amounts to stating two generalizations and adding an exception for the underapplication. The general mechanics of this solution would not work for overapplication due to counterbleeding, but there is always provably some set of constraints, however unnatural, that will capture the pattern (Kaplan & Kay 1994). For example, in this case, the raising rule overapplies. Clearly we cannot ban raised vowels before voiced segments, because the flap is voiced, but we can ban raised vowels before each of the voiced segments except for flap; the trick is then constructing constraints that handle pattern for flapping environments, which is to raise only preceding an *underlying* voiceless segment. But constructing such constraints is not a problem for monostratal Optimality Theory: it is a problem for a version of correspondence theory that only allows statements about dependencies between input and output to be made with respect to a surface segment and its single input segment. What is needed here is dependent on the *next* input segment; but, again, this is not an issue for the architecture, it is a question about the theory of Con.

ule of the phonology, as in the current architecture, or as is generally the case in Lexical Phonology (Kiparsky 1985),<sup>4</sup> then its output is not represented at the surface (of the other module), and so the problem of getting the underlying representation of an allophonically obscured environment to be visible dissolves: there is no change at the surface representation. The problem becomes slightly more complex if, as here, the process whose environment *is* obscured is also arguably strictly allophonic—and thus also in that second module itself, and thus potentially re-obscured by the effect of the second process applying in that module. The current architecture, however, solves this handily, because, unlike in standard Lexical Phonology, the allophonic processes apply as if simultaneously (see above and Chapter 5).

We have thus raised the following idea: certain opacity puzzles in surface-oriented phonology rest, naturally, on a particular conception of the surface which is not the one we are discussing here, just like the analysis of post-velar fronting and many others; but, more than this, *the very fact that they are puzzles* rests on this particular conception of the surface. Once a different architecture is adopted, the problem of visibility of environments on the surface evaporates for processes in the second module; and given the character of that second module, as proposed here—input-oriented, but not derivational—the problem of opaque interactions evaporates within that module, not in spite of, but *because* of the

---

<sup>4</sup>One can easily list the bevy of alternate solutions proposed to replicate these invisibility effects, which are given rise to by the inherent ordering of the post-cyclic/word-level or post-lexical levels in Lexical Phonology. A case of a post-cyclic rule that has received a lot of attention is the Levantine Arabic epenthesis rule discussed by Brame 1974. A short epenthetic vowel is inserted in certain environments, but it seems to be invisible to the phonology. See Kiparsky 2000 for a survey of recent solutions to this particular problem that attempt to replicate the invisibility effect without importing the architectural cut. See below for further discussion of the relation to Lexical Phonology. I will not discuss this case further, not because I do not believe that the epenthesis is plausibly gradient, but because attempting to develop a theory of phonetic processes that can handle wholesale insertion would take me too far afield.

fact that the phonetic module is not derivational. Again, this will be taken up in greater depth in Chapter 5.

In this section, I have gone through three examples of analyses that make explicit reference to SC-representations: in one, it was not crucial; in another, it was crucial, and we developed a new analysis of these historical facts which fixes the problems for the SC-crucial grammar, and removes an implausible-sounding leap from one grammar to another from both analyses; and, finally, we went through a case where representing allophony as phonetic made a very big difference to how the grammar could be stated, namely, a particular case of opacity.

There are, of course, many other simpler cases of reliance on SC-representations which do not require much elaboration: to reiterate, all uses of “surface phonotactics” to capture allophony—whether the case is one, like Mielke et al.’s analysis of Canadian English, where the result is supposed to be wholly static knowledge, with no alternations, or one where the surface phonotactics is supposed to give rise to morphological alternations as well—all rest on crucially incompatible assumptions. This was discussed in the context of learning models in Chapter 3, a list which includes, minimally, Boersma 2001, Boersma & Pater 2007, Hayes & Wilson 2008, Blanchard & Heinz 2008, Jarosz 2011, Elsner, Goldwater & Eisenstein 2012. Although these learners generally do quite well, Chapter 3 argued that, in the face of the difficulty in learning SC-representations, the (relatively simple) task of learning allophony was put to more practical use if embedded in the phonetic learner, allowing for a reduction in the complexity of the surface representation, from an SC-representation to an AC-representation, taking away the motivation for applying learners like this to allophony. This section has outlined the broader framework

in which this is situated.

In the next section, I will probe the consequences more deeply, highlighting the connection to Lexical Phonology (Kiparsky 1982).

## 4.2 The Lateness of Allophony

### 4.2.1 Background: Structure-preservation and the cycle

Since the linguistic computation needs to map not just between phonetics and *strings* of lexical items, but in fact between phonetics and *hierarchically organized* collections of lexical items—morphosyntactically complex objects—something needs to be said about how the phonological computation interacts with this structure; one approach would be to assume that the phonological component just maps between phonetics and an unbroken concatenation of all the lexical items in an utterance, completely insensitive to any morphosyntactic structure. Another approach would be to make the phonology sensitive to only the lowest-level boundaries—divisions between the leaf nodes of the morphosyntactic tree, but nothing more. However, another approach suggests that the phonological computation hews closely to the morphosyntactic tree. The *phonological cycle* was introduced in the 1960s as a strong instantiation of this idea. In the original version, the phonological grammar applies once to generate an output string at the leaf nodes; then erases the first layer of bracketing and applies again; and then erases the next layer of bracketing and applies again; and so on, until all brackets have been erased. This did a fair bit of work in the *SPE* analysis of English stress, so that the derivations for *orchestration* and *infestation* were, roughly:

		$[[rkstret]_Vjn]_N$	$[[nfst]_Vetjn]_N$
First cycle	Main stress rule	$[[rkstret^1]_Vjn]_N$	$[[nf^1st]_Vetjn]_N$
	Alternating stress rule	$[[^1rkstret^2]_Vjn]_N$	–
	Pretonic weakening	–	–
	Auxiliary reduction rule	–	–
	Spirantization	–	–
	Palatalization	–	–
	Glide deletion	–	–
	Bracket erasure	$[^1rkstret^2jn]_N$	$[nf^1stetjn]_N$
Second cycle	Main stress rule	$[^2rkstret^1jn]_N$	$[nf^2stet^1jn]_N$
	Alternating stress rule	–	–
	Pretonic weakening	–	$[nf^3stet^1jn]_N$
	Auxiliary reduction rule	–	$[^2nf^3stet^1jn]_N$
	Spirantization	$[^2rkstres^1jn]_N$	$[^2nf^3stes^1jn]_N$
	Palatalization	$[^2rkstre^1jn]_N$	$[^2nf^3ste^1jn]_N$
	Glide deletion	$[^2rkstren^1]_N$	$[^2nf^3sten^1]_N$
	Bracket erasure	$^2rkstren^1$	$^2nf^3sten^1$

Then, after the cyclic rules applied, a *post-cyclic rule* of vowel reduction applied in syllables not marked with stress:  $^2rkəstre^1ən$ , but  $^2nf^3ste^1ən$ . The morphological differences in the two words—deriving from *orchestrate* versus *infest*—combined with the cyclic application of the stress rules to protect the second syllable of *infestation* from reduction. The emergent descriptive generalization about English stress was later called into ques-

tion, and a new analysis, which did not make use of cyclicity in this way, was proposed (Halle & Kenstowicz 1991: see below for discussion of the other implications their analysis had). However, the inside-to-outside application model had in the meantime accrued several other empirical patterns to its credit; one type, *Strict Cyclicity Condition*, or SCC, effects, which were handled under a small modification to the Chomsky and Halle framework; and *level ordering* effects, which prompted a more radical revision to the model in the form of Lexical Phonology. The latter will be discussed shortly.

In all of the cyclic computation models, however, a second block of non-cyclic rules had to always be distinguished. For example, vowel reduction had to apply once, *after* the cyclic computation; it had to be after, or else the stress placement rules would not have worked properly. In the grammars of many other languages, such post-cyclic rules also appeared. A curious “post-cyclic syndrome” then began to emerge: post-cyclic rules were *non-structure preserving*, which meant that they almost never hewed to the (hypothesized) underlying inventory of the language; that is, they were, largely, strictly allophonic. This cut in the phonology did not have anything to do with the cycle per se. Even without the cycle, a division of phonology into early structure-preserving rules and late allophonic rules was plain: having established a partial order over the rules in a grammar, one would almost always find that no non-structure-preserving rule was ordered before a structure-preserving rule (in fact, as I will discuss in Chapter 5, within the late block, the rules also almost never fed each other). A division saying roughly that structure-preserving rules fed non-structure-preserving rules was not new. In fact, one type of pre-generative theory that had a division like this had driven one of the early vociferous battles surrounding generative grammar, the debate over and break from the structuralist taxonomic phoneme

level. However, this only became clear within generative grammar once there was an architectural cut that could make the distinction clear. The other important symptom of the post-cyclic syndrome was that post-cyclic rules were without lexical exceptions, whereas cyclic rules were generally riddled with them.

To fully understand the context in which we should be talking about allophony and lateness, we must first cover a few other facts and theories. The relevant theories all incorporate hierarchically structured *cyclic domains* for the application of phonological grammar. In the *Aspects/SPE* model, one could see a “domain” as corresponding to a subtree of the morphosyntactic structure assigned to the sentence. This model became less popular after an effort was made to separate morphological from syntactic structure in different components of grammar. At the same time, within the morphological system, a demand for higher-order cyclic domains emerged. *Lexical Phonology* maintained the cycle, but divided it up into different strata, or *levels*. The level L(I) was for all the morphology closest to the stem; in English, it was home to the stress-shifting affixes (-ity, -ic, -ate, ...); the level L(II) contained stress-neutral affixes (agentive -er, -ness, -hood, ...), as well as compounding. This cleaned up a distinction that was already made in *SPE*, wherein what wound up being the L(I) affixes were treated as lexical-item-internal boundaries marked with a “formative boundary symbol” +, a weaker boundary than the “word boundary symbol” # introduced between layers of morphology. Further levels were also proposed, with authors disagreeing on how many there should be (the current consensus within Stratal OT, the modern descendant, is two). Each level could in principle have its own set of rules, although some authors posited constraints that forced some sharing across levels (Halle & Mohanan 1985). The principles of the cycle (most notably the the

recently discovered SCC) were generally understood to still hold *within* levels, however; thus there were two types of domain: the morphological cycles, and the levels that set off a nested grouping of these morphological cycles, with L(I) affixes nested within L(II) affixes and so on, but still giving rise to cyclic domains within levels. More recent work, mostly following Distributed Morphology, (Halle & Marantz 1993), has returned to the integrated morphology–syntax model. In the meantime, syntax, too, has developed a type of higher-order cyclic domain than just the bare syntactic subtree. There have been several versions since the 1970s, (Bresnan 1972, Uriagereka 1999, Chomsky 2001), and in most current versions these higher-order cyclic domains—specially designated subtrees, CP, vP, DP, and sometimes others—are called *phases*. Recent work has attempted to unify the phonological cyclic domains with Phase Theory (Piggott & Newell 2006, Compton & Pittman 2010, Slavin 2012). Still other types of theories have suggested domains defined by an independent phonological (prosodic) hierarchy, unrelated to the morphosyntactic structure, but these are largely irrelevant for current purposes.

In this section I will discuss two things: first, the deduction of the order *non-structure-preserving follows structure-preserving* under the current architecture. Then, since other work on this question is usually embedded in a model with cyclic domains, (either just in the *Aspects/SPE* sense or also in the second-order Lexical Phonology/Phase Theory sense), I will discuss how this architecture comports with those theories. The first part will be easy, the second part will be left open as clearly problematic. That is because, as we will see, a natural way to interpret the allophonic rules when we combine the current architecture with cyclic phonological domain theory will be as rules applying at the end of *every* cyclic domain. This is natural either because (i) such a thing already has some em-



pirical support, with blocks of post-cyclic rules applying at the end of every higher-order domain (Lexical Phonology); or (ii) because the theory demands that the higher-order domains be domains for phonetic interpretation (Phase Theory), and making this empirically meaningful demands that we try to hew as close as possible to having this mean actual phonetic interpretation in the gradient sense we have been working with. This is problematic. I outline the reasons and sketch some suggestions for proceeding at the end.

To clarify point (i): Bermúdez-Otero & McMahon 2006 and Bermúdez-Otero 2013 attribute the following examples of blocking of allophonic processes in English to cyclic domain differences:

(101) *Vernacular London Goat Split*

→ /—l. where . represents a syllable boundary

	Stem	L(I)	L(II)
<i>hole</i>	hl		<i>holey</i> hl#i
<i>holy</i>	h.li	<i>Walpolian</i> wl.p.l + iən	

(102) *Northern Irish Dentalization*

{t, d, n, l} → dental/—(V)r

Stem	L(I)	L(II)
<i>train</i> tren	<i>sanitary</i> sænt + æri	<i>shouter</i> tʃəʊr

(103) *Canadian Raising*

{a, ʌ} → central, raised/  $\left[ \begin{array}{c} - \\ \langle \text{stress} < 1 \rangle \end{array} \right] \left[ \begin{array}{c} \text{C} \\ -\text{voice} \end{array} \right] \left[ \begin{array}{c} \text{V} \\ \langle -\text{stress} \rangle \end{array} \right]$

Stem	L(I)	L(II)
<i>Eiffel</i> fəl	<i>i-th</i> +θ	<i>eyeful</i> a#fəl

The generalization they give is that L(I) affixes form part of the domain for allophony—the end-of-domain rules apply at the end of L(I). (In fact, as we will see below, if this generalization is correct, then it is a better diagnostic for the cyclic domain of an English affix than the usual diagnostic of stress-shifting, because that has since been undermined.) Thus one might think that the appropriate sense of “late” is “at the end of L(I),” at least for English. However, there are also allophonic processes that apply across both L(II) suffix boundaries and indeed word and phrase boundaries. The first set of allophonic processes are usually called “post-cyclic” and the second “post-lexical” (Booij & Rubach 1987). This is extremely confusing, and so, to keep things straight, I will continue to refer to blocks L(I) and L(II) as particular examples of “higher-order cyclic domains,” which I

now begin to abbreviate as HOCDs; the non-cyclic rules applying at the end of each block become, in my terms, *end-of-domain rules*, EOD rules. If this model is right, then there are separate blocks of allophonic rules for each HOCD; if it is not right, then there is only one block of allophonic rules, following the outermost HOCD (whether there are HOCDs or not). I begin, however, by putting this aside—just assuming that there is one block of allophonic rules—and deducing their lateness.

#### 4.2.2 Structure-preservation and phonetic transforms

“Deducing” the lateness of non-structure-preserving processes may sound strange given the architecture proposed above. Particularly if we continue with our simplifying assumption, made up to this point, that the phonetic transforms *must* be non-structure-preserving and the categorical phonological grammar *must* be structure-preserving—which we will elaborate on and nuance here and in Chapter 5—it sounds like, by proposing this architecture we have already just stipulated that non-structure-preserving processes follow structure-preserving processes.

The point is this: even before we ask about the association between phonetic interpretation and non-structure-preservation, we can ask the separate question of whether the order “categorical phonological grammar before phonetic interpretation” *must* be the architecture, and indeed it must. Consider the following very weak core principle of what we understand to be meant by *phonological* grammar—what I have been distinguishing from phonetics by calling it *categorical*. What makes it “categorical” is, minimally, this:

(104) **Coarseness of Phonology Principle.** The set of distinct single-segment

representations which appear in the set of all possible grammatical outputs for a given phonological grammar can be placed in correspondence with only a proper subset of the universally possible single-segment phonetic representations.

This says that, for any particular phonological grammar, the set of segmental distinctions made in legitimate grammatical outputs, in toto, is of lesser cardinality in the Cantor sense than the set of segments that the phonetic system can distinguish. It does not exactly say that the set of universally possible segmental phonological representations is of lesser cardinality than the set of universally possible segmental phonetic representations: but, in that case, a grammar that placed no constraint, either input or output, on the grammatically possible segments, so that there would therefore be a way of phonetically interpreting the outputs that allowed for any possible phonetic distinction, would violate the CPP. Such a grammar would leave us without any explanation for the basic empirical fact discussed in Chapters 1 and 3 that the changes phonological grammars make are categorical.

Now, suppose that, under our model, we fix a context  $x$  and execute the mapping from possible phonological output segments to their corresponding phonetic interpretations  $\mathcal{P}_x$ . For any one of the possible segments, suppose we then change the context to  $x'$  and find the result  $p_{x'}$ . There is nothing in this architecture that says that  $p_{x'} \in \mathcal{P}_x$  or any  $\mathcal{P}_{y \neq x'}$ ; only assumptions about phonological grammar that violate the CPP could ensure such a thing. Therefore, even if there were some segmental category that represents  $p_{x'}$  in another context, (think: some way of representing Inuktitut [o] in the phonological representation other than the sequence [u][+uvular]), logically, the phonological grammar would be forced to neutralize it to something else; anything else would be impossible by

hypothesis. For all intents and purposes, this means that it is perfectly legitimate for there to be some (oracle) category in this system for which the only possible way of generating it is by applying a transform to a particular segment in one particular context.

Now it is a short jump to the conclusion that context-dependent phonetic interpretation cannot feed phonological grammar. For if it did, its application would sometimes be uninterpretable or fruitless, restored to some new categorical representation only to be inevitably neutralized to some other.

Notice that this does not imply that all the segments in the input need to be phonetically interpreted at the same time. It is fully consistent with this feeding constraint on the architecture that we could have a variant architecture, tolerating “mixed” interpreted–uninterpreted sequences. Above, we happened to stipulate this out of existence, but were it true, the reasoning here would tell us that, during the phonological computation, there could be already-interpreted elements that the phonological grammar would have to studiously avoid ever making reference to. This would be almost indistinguishable from the architecture proposed here, except that (i) it would open up the new possibility of gradient environments for allophones; and (ii) it would allow, in principle, for cyclic non-structure-preserving processes, diagnosable perhaps by their otherwise unexplained restriction to derived environments (SCC; see Kiparsky 1982). As far as I know, there are no such cases, and so, for the time being, point (ii) remains moot, and point (i) remains barred by stipulation. It also does not rule out the more general possibility that one could have phonetic interpretation feed phonological grammar, but then, only those representations which could be restored as phonologically interpretable categories could be operated over further. The same reasoning applies: this is basically indistinguishable, although it would

be worth investigating certain exotic cases further.

Having shown, then, that phonological grammar more or less has to feed phonetic grammar, and not the other way around, I now consider what this has to do with non-structure-preservation. Recall that the fact to be explained is that non-structure-preserving rules are late; the obvious late block in our architecture being the phonetic component, it is natural to try to show that category changes which are non-structure-preserving in the classical sense—introducing a distinction made nowhere in the lexicon, or explicitly barred in the lexicon—are phonetic.

Now: what I have just outlined amounts to saying that phonetic transforms are potentially non-structure-preserving in a different sense—they are at least capable of introducing distinctions banned in the output, not the input, to the phonological grammar. I will show how we get from here to an implication in the correct direction; and then to an implication, not from “non-structure-preserving in the classical sense” but from “in complementary distribution.” Now, throughout this chapter and the preceding one, I have made numerous references to the idea that the association between phonetic transforms and allophony was not a perfect one. Thus what I will show here is simply how to reverse the implication, given that it is imperfect in both directions (the details will be saved for Chapter 5). The answer, of course, is Bayes’ Rule.

Now, there are two different claims being made: the first is about non-structure-preservation in the classical (lexical inventory) sense, and the second is about complementary distribution. I will start with the first one, about processes giving rise to non-lexically-specified categories. We start with this fact which will hold of any gradient phonetic learner:

(105) The prior probability of the (context-transformed) phonetic representation for any category being exactly identical to any the (context-transformed) phonetic representation for any existing category is extremely low. Let  $o_{i,x}$  be the oracle category corresponding to category  $i$  in context  $x$ . For any category–context pair  $j, x' \neq i, x$ :  $\Pr[o_{i,x} = o_{j,x'} | \text{lexicon, grammar}] \ll \Pr[o_{i,x} \neq o_{j,x'} | \text{lexicon, grammar}]$ .

Now, hold fixed a lexicon in which there is a collection of tokens that are all coded using the same category, even though any high-likelihood phonetic implementation will split them into two different oracle categories. Bayes' Rule can help us to go from the above to see just when it will be the case that  $\Pr[\text{transform } i \text{ in ctxt } x | \text{lexicon}, \forall j, x', o_{i,x} \neq o_{j,x'}] > \Pr[\text{new cat. in ctxt } x | \text{lexicon}, \forall j, x', o_{i,x} \neq o_{j,x'}]$ . That is, when will it be the case that an oracle category in some context being distinct from any existing oracle category given by the lexicon and the existing phonetic implementation (thus, non-structure-preservation) means a phonetic transform (late rule) *is better for the learner* than a phonological rule?

I will spell out all the technical and empirically testable details in Chapter 5.

It will, of course, have something to do with complementary distribution. To see this, think about the following relevant facts, which speak to the converse of the second claim:

(106) Given a grammar with phonetic transforms perturbing category  $i$  in context  $x$ , the probability of there existing in the phonetic interpreted output a distinct oracle category for context  $x$  is just the probability of the phonetic transform for  $x$  being non-zero:  $\Pr[\text{transform of } i \text{ in context } x \neq 0 | \text{lexicon, grammar}]$ . Given that the transform is non-zero, there will be two oracle categories in complementary

distribution.

Bayes' Rule can help us go from the above to see just when it will be the case that  $\Pr[\text{transform } i \text{ in ctxt } x | \text{complementary distribution}] > \Pr[\text{new cat. in ctxt } x | \text{complementary distribution}]$ . Suffice it to say, this will generally be true. I leave the details, as well as the discussion of purported cases where this preference seems *not* to have held, despite complementary distribution, for the discussion of incomplete neutralization in Chapter 5.

### 4.2.3 Issues with phonetic EOD blocks in HOCD theories

Consider now the empirical evidence given by Bermúdez-Otero and McMahon that led us to conclude that English post-cyclic rules applied within a narrow HOCD we called L(I); this domain is well within the scope of primary stress and resyllabification, setting up a bifurcation in the domain corresponding to firm “word”-like boundaries. As the notion “word” is neither a surface property nor obviously a natural kind, this is unsurprising. However, if HOCD blocks all have phonetic implementation at the end of them, then it implies phonetic implementation in some sense “within words.” Crucially, if the inside-to-outside computation model is correct, then the mapping to phonetic output for a nested HOCD is a sufficient substitute for the tree in the mapping at the outer HOCD. To the extent that phonetic interpretation in the “outside” domain depends crucially on phonological, rather than phonetic, information “inside” a nested domain—call this *illegal HOCD interaction*—then one of two things needs to be done: (i) we need to claim that phonetic interpretation can actually be undone; (ii) we need to explain why that information also needs to be present in the phonetic object anyway. At least where the relevant



phonological information is some categorical representation of an allophone, (i) should surely be impossible, but any case where we need to appeal to that device weakens the notion of “phonetic interpretation” substantially.

I now go through some cases that look problematic on the surface. The first type is stress adjustment. The positions of phrase and compound stresses are almost always the relevant stresses as determined word-internally—that is to say, the position of stress is computed at the inner domain and not adjusted at any outer domain—but the final *degree* of stress is in general determined at the outer domain (see Chomsky & Halle 1968 for English nuclear stress, and see Cinque 1993 for German and Italian; although some have suggested that the position of secondary stresses within German words is mobile and depends on sentence context, Knaus, Wiese & Domahs 2011 present clear evidence that it is determined within the same domain as main stress). The degree of stress is gradient information; however, it is hardly “phonetically interpreted.” To the extent that sentence-level stress adjustments preserve the language-specific phonetic interpretation of the abstract “degree of stress” marking, it must remain abstract, subject to the latest possible interpretation. Thus, posting end-of-domain phonetic interpretation commits us to a two-stage process of phonetic interpretation for stress; it also seems to commit us to saying that determining sentence-level degree of stress is done in a way that is qualitatively different from the way word-level degree of stress is marked, which is inconsistent with many metrical theories.

Altering the degree of stress on a particular syllable, whether the change is gradient or not, requires identifying the position of the stress-marked syllable in some representation. The position of stress is treated as a position in a sequence, on a grid line, or in a tree;

but at any rate it is discrete. One might, at first blush, think that this could explain why the position of word stress is not changed at the phrasal level: we might expect that the discreteness of the sequence is erased. This would not make sense, however, as the correct syllable still needs to be identified in order to change the degree of stress. Indeed, there is no particular reason to think that the structure scaffolding the position of stresses is ever required to be converted to a representation which is gradient in the relevant sense. A few cases where the position of word stress is apparently adjusted according to position in a larger domain do exist, and they are informative. The principle case is stress retraction, as in the English rule shifting *Marcèl Próust* to *Màrcel Próust* and the related German as well as Tiberian Hebrew patterns (see Prince 1975, Liberman & Prince 1977, Dresher 2009a); analyses of these retraction phenomena using a hierarchical metrical grid or tree only require moving the top-level prominence marking to align with a position fully determined by the next level of the structure. Thus the visibility of this information could yet be heavily restricted. Even granted that phonetic rules can access metrical structure, however, one wholly problematic case is posed by the pausal forms in Tiberian Hebrew (Prince 1975). These marked forms crucially appear with a restricted distribution—before a major phrase boundary:

UR		Contextual	Pausal
(107)	katab+ta <i>you wrote</i>	katabta	katabta
	katab+u <i>they wrote</i>	katbu	katabu
	zaqen+u <i>they are old</i>	zaqnu	zaqenu

A recent analysis by Dresher (2009b) makes the computation of word-internal metrical structure sensitive to the prosodic context, nearly from the start: immediately after the

initial unmarked main stress rule applies, (penultimate for vowel-final words, otherwise final), a strengthening which affects metrical structure takes place in the pausal-form context: in Dresher's metrical grid account, in the pausal-form environment, "assign a right bracket and a grid mark at every prosodic level up to the Intonational Phrase on a vowel to the right of a left bracket"; vowels which head an Intonational Phrase are then lengthened. Since there are no subsequent metrical operations needed at the lower or intermediate levels, this derivation itself is unproblematic, assuming that the lengthening is phonetic at the IP level; but the environment for this strengthening is inherently phrasal, requiring the "inner-level" computation to be sensitive to a small amount of "outer-level" information. At present, this is a case of strictly illegal HOCD interaction; both this exceptional case and the fact that it is exceptional must be explained.

One might think, however, that stress is just different and is not subject to the bounds of the HOCD architecture. Segmental phonology presents a much bigger problem. Generally speaking, any segmental rule which applies across "word" boundaries is going to be a problem if "word" is understood to refer to some HOCD triggering phonetic interpretation. The Polish rule of allophonic palatalization (Booij & Rubach 1987) affects all consonants before [i] and [j]. It also applies across word boundaries. That means that the information "consonant" and "high front vowel" needs to be preserved in the output of phonetic interpretation. It is possible to reconcile this with a weak sense of phonetic interpretation: the segment [i] is spelled out, but this information can be converted back to categorical information if necessary. Contrast this with our architectural argument above: there it was the fact that allophones could not be mapped back to their own phonological categories that blocked further interactions; but the relevant information here is not

introduced by allophony. (For now I am putting aside the problem raised by Booij and Rubach's claim that a word-level EOD rule of retraction bleeds palatalization; see Chapter 5.)

Similarly, an inner HOCD may provide the trigger, not the target, for a process in an outer domain. Kiparsky 1982 and Halle & Mohanan 1985 assign English regular inflectional morphology (plural  $[-z]$  and past tense  $[-d]$ ) to L(III) and L(IV) respectively, for simple reasons of its position outside derivational morphology and compounds. The standard analysis is that both are subject to epenthesis of a short vowel following stem-final sibilants (plural) or coronal stops (past); then devoicing following stem-final voiceless obstruents. Both processes clearly depend on phonological information in the stem which appears to be categorical.<sup>5</sup> Again, the only way to handle this is to weaken the notion of "phonetic interpretation." (It is also worth noting that the standard analysis, with two separate processes in a feeding relation, could not work here if the processes are both phonetic: see Chapter 5.)

As for stress, it is usually accepted that there are no illegal metrical interactions of the kind discussed above within English words: one of the key properties L(II) affixes is that, unlike L(I) affixes, they are never stress-shifting (although *-al* is stress-sensitive: *arrival*, \**edit-al*). However, as Halle & Kenstowicz 1991 point out, stress-neutral L(II) affixes in fact sometimes mysteriously appear strictly inside the scope of stress-shifting L(I) affixes, and not only receive stress as a result, but appear to do so over their underlying,

---

<sup>5</sup>English homorganic nasal place assimilation is another such possible case (*un-* is an L(II) affix; *un+believable* assimilates in place to  $[mb\acute{o}liv\acute{e}b\acute{e}l]$ ); this is likely an EOD rule, as it also generates *un+fair*  $[fr]$ , with a result that is not lexically contrastive. However, the spreading nature of this interaction suggests another analysis in which the triggering information really is phonetic. If that means gradient, then we could in principle predict gradient effects depending on small deviations in the pronunciation of the following place—so, more like coarticulation—but in fact this reasoning is highly debatable.

pre-interpreted segmental representations: *pátent#abəl* versus *pàtent?abíl+ity*. Halle and Kenstowicz simply mark each suffix as being stress-shifting or not; each stress-shifting affix completely erases the stresses previously assigned and starts the metrical structure anew. This undermines the assignment of these suffixes to separate HOCDs, but there is no reason to think that stress needs to be computed at the innermost HOCD.

Contrast this with the Latin clitic–host stress pattern discussed by Halle & Kenstowicz 1991. Stress is antepenultimate if the penult is light. Stress can shift under enclisis (*úbi*, “where,” *ubí#libet*, “wherever”), but it does not always shift to the position predicted under the stress rule (*límina*, “doorways,” *liminá#que*, “and doorways,” \**limína#que*). The explanation does not rely on *-que* destructively modifying the existing metrical structure; rather, it relies on the fact that it cannot. The addition of *-que* assigns a line 1 asterisk, but only to unbracketed line 0 asterisks—not to already bracketed and projected ones. The first two syllables of *limina* already having been bracketed, the only legal place to mark an asterisk is at the third syllable. No destructive modification of information in the nested HOCD is needed to get the stress shift, suggesting that in Latin it would be possible to compute the stress at both HOCDs, and limit interactions in the way suggested above. (For Halle and Kenstowicz, it does require a revocation of extrametricality on the last syllable, but this is an idiosyncrasy of their theory.)

Finally, resyllabification phenomena are pervasive and clearly require access to categorical information, but only up to a point: (*sea*[l]), ((*sea*.[l])(*o.ffice*)), where compounding appears to be L(II), but ((*the*)(*sea*[l]))((*offered*))((*a*)(*donut*))), with no resyllabification at a higher HOCD. The categorical identity of [l] is necessary in order to recompute the light/dark allophony.

In sum: the well-known association between non-structure-preservation and late application (with its notable consequence of failure to be blocked in non-derived environments) is deduced under the current architecture, but needs to be weakened in an inside-to-outside phonetic interpretation theory. In particular, although the crucial step in the deduction makes use of the fact that only gradient information is visible after interpretation—and gradient interpretation is incompatible with categorical phonological grammar—any attempt to compute the phonetic interpretation of an entire utterance by doing complete interpretation of its parts seems to run into immediate difficulty by this logic, because computation seems to require some access to categorical, and not only gradient phonetically interpreted, material inside nested domains. I noted that some limited amount of interpretation-reversal could alleviate the problems. See Chapter 5 for a bit more discussion.

Finally, I note that some measure of constrained non-structure-preservation in the cyclic phonology can be tolerated; it is not a hard constraint. On the other hand, as first hinted at by Kiparsky 1985, any cases where post-cyclic processes are genuinely structure-preserving substantially weaken a theory of gradient post-cyclic processes, and that is certainly also true here. (A relevant case brought up by Kiparsky is Vata ATR harmony, but his question—he does not know the answer—is not whether it is structure-preserving, but whether, given that it is not, it bears Cohn 1990’s interpolation-type hallmarks of phonetic harmony. This would be an interesting case to examine further phonetically.)

### 4.3 Summary

In this chapter, I have addressed two consequences of taking seriously the notion of context-dependent phonetic interpretation, as proposed in the current architecture: under the assumptions of LPTH, I have discussed the absence of conventional “surface representations” coding allophones in a categorical way — which I call SC-representations, to suggest “surface category” — and their replacement with more abstract representations, which do not code predictable allophones — which I call AC-representations, to suggest “abstract category.” I have also discussed non-structure-preservation, the definitional property of strict allophony, and deduced the “phonology-first” architecture, which is this theory’s answer to the “post-cyclic” or “post-lexical” rule block in Lexical Phonology and related theories. I have pointed out some of the problems with this architecture, when taken seriously as a model for cyclic phonetic interpretation, and suggested tentative solutions.

## Chapter 5: Phonetic transforms II: Linguistic phenomena

Everything that happens will happen today,  
and nothing has changed, but nothing's the same.

—David Byrne, “Everything That Happens Will Happen Today”

In this final chapter, I show that incomplete neutralization is not only easy to handle on the current theory, but actually predicted to be a very general case for phonetic rules. I show how the posterior evaluation measure handles various key phonetic patterns (strict allophony, incomplete neutralization, true neutralization), and why the first should tend to be phonetic, the last should tend to be phonological, and why incomplete neutralization might sometimes be learned as complete (phonological) neutralization, leading to language change. I then show that the opaque ordering exhibited by Canadian Raising is not only easy to handle, but predicted to be the only possible type of interaction between phonetic transforms. In order to maintain this empirically, I need to clearly separate the kinds of phonetic processes handled by phonetic transforms from another type of apparently phonetic process, phonetic spreading, which I propose takes place via (deletion and) sharing of phonetic features. I finish by discussing a few outstanding issues.



## 5.1 Incomplete neutralization

Nothing in the theory above says anything about how phonetic transforms deal with the existence of other phonetic categories in the system. Although we have alluded to the idea that phonetic transforms create “new categories,” not generally expected to be identical to the realization of any other existing phonetic category, we have not said anything about the possibility that the result might be very, very similar.

*Incomplete neutralization* is exactly this. Incomplete neutralization refers to a process that gives contextual variant pronunciations that are similar, but not phonetically identical, to some other segmental category. The best known example of incomplete neutralization is German word-final devoicing (Port & O’Dell 1986); however, many other voicing alternations show evidence of being incomplete (Catalan: Wheeler 2005; Dutch: Ernestus & Baayen 2006; Polish: Slowiaczek & Dinnsen 1985; Russian: Pye 1986). The English flapping alternation is arguably incompletely neutralized for voicing, a fact which will be discussed later in this chapter (Braver 2011). Turkish devoicing does not appear to be incomplete, on the other hand, although the data (Wilson 2003) is not perfectly clear, and the Korean neutralization of  $t^h$ ,  $s$ ,  $s^*$ ,  $t$  and  $t^h$  to  $t$  in coda position is also evidently phonetically complete (Kim & Jongman 1996).

To give a clearer picture of the phonetic effect of incomplete neutralization, I have combined the results of two acoustic studies on German stops in Figure 5.1.

Figure 5.1 shows that stops in both final and initial position are less phonetically different than stops in intervocalic position (the means are closer together).<sup>1</sup> Stops in ini-

---

<sup>1</sup>The intervocalic position I have chosen rather arbitrarily as the basis for comparison just because it

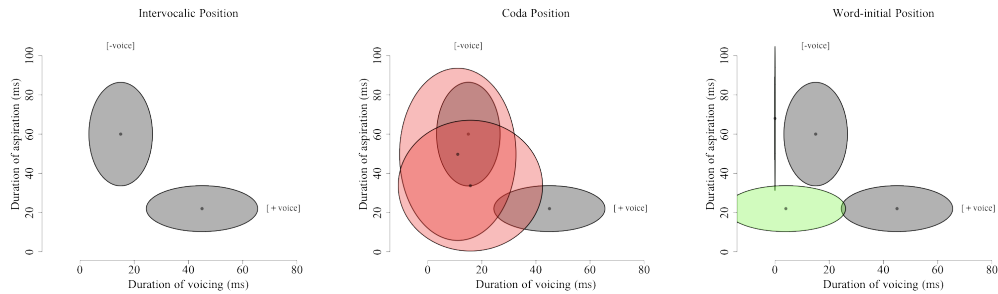


Figure 5.1: Duration of voicing versus duration of aspiration for German stops in three positions. The left panel shows intervocalic position (a “lenition” position); the middle panel superimposes coda position (the “final devoicing” or “final fortition” position); and the right panel superimposes word-initial position (the “initial fortition” position). The data are taken from summary statistics presented by Jessen (1998), except for the coda position data, which are taken from summary statistics presented by Port & O’Dell (1986). Categories are plotted as though they were Gaussian, even though their distributions are all truncated at zero. Categories are plotted as though they had zero covariance, but their covariances were actually just not reported in either study.

tial position show little to no detectable voicing (in fact always zero for voiceless stops) but the aspiration contrast is roughly the same. Stops in final position show an incomplete neutralization in both voicing and aspiration, but their voicing remains phonetically distinct from initial position (almost certainly because they are preceded by sonorants in final position; for final and intervocalic position the voicing is so-called “voicing into closure,” measured from the offset of the preceding sonorant, while for initial position any voicing prior to aspiration is counted, as the point at which closure is initiated cannot be determined from the phonetic data). The data is presented principally for illustrative purposes. As far as I know, no single experimental study has measured both the final devoicing context and either of the two contexts studied by Jessen. As the experimental conditions and the tools used for measurement were somewhat different across the two studies, the comparison highlights the contrast best. This environment is usually understood to be subject to “additional” voicing, but whether one of these environments actually represents the basic phonetic representation of the contrast, and, if so, which, is impossible to know just by looking at this data.

parison of the results is imperfect. The main point, however, is that, in both cases, the laryngeal contrast is partly neutralized, but underlying voiced and voiceless stops remain phonetically distinct.

Much of the debate surrounding incomplete neutralization during the 1980s seemed to rule out the possibility that it could coexist with true neutralization: the followup study of Fourakis & Iverson 1984 tried to rule out the possibility that speakers were enhancing the contrast on the basis of the orthography, on the assumption that anything else would be problematic; and Port & Leary 2005 explicitly uses incomplete neutralization as an argument against the existence of categorical phonological processes. However, the current architecture is in line with a handful of recent proposals that recognize the need for both categorical and gradient context-dependent operations (van Oostendorp 2006; Braver 2011). Furthermore, given the discussion of Old East Slavic post-velar fronting in Chapter 4, the same grammatical mechanism may be assumed to be at work in phonetic enhancement, and so both things—that voicing can be incompletely neutralized, and that the degree of neutralization can change as a response to pressure to enhance a contrast—can be true.

The question, of course, is how incomplete neutralization, or more generally, phonetic grammar, can coexist with phonological grammar, **and still permit phonological grammar to do its job**. The basic questions one needs to answer, then, are what evidence would lead the learner to think that it should be attributable to a phonetic rule and which would not. In Chapter 4, we said something to the effect that complementary distribution would lead a learner to posit a phonetic rule; in this section we flesh that out. Complementary distribution will of course break down to a fair degree for a real speech perceiver

(not an oracle) under incomplete neutralization; I discuss when the learner would find it appropriate to maintain a phonetic rule nevertheless, and when it will abandon this. I then discuss briefly the kinds of additional pressures that will lead a learner to alter the phonological grammar to explain the effect as a “real” neutralization.

Before moving on to the main discussion of how these patterns will be treated by the learner, however, it is worth noting some general patterns that seem to hold in the phonetic data. The order holding between the means across categories on both dimensions remains fixed across all three contexts, but the difference is smaller in the fortition positions, except for initial position, where the difference in aspiration is slightly larger than in intervocalic position. The ordering between the variances along the two dimensions also remains fixed within categories and across categories, across all three contexts; in initial and intervocalic position the variances are the same for voiced stops. Future work should attempt to check these descriptive generalizations to rule out the possibility of interference from the difference in methodology (the numerical difference in the variances in coda position is particularly salient, but in principle any or none of these generalizations could be spurious). They are all suggestive of the pattern laid out for Inukitut and Kalaallisut retraction and fronting in Chapter 3, and further study will help to support generalizations about  $\oplus$ . Regardless of the details of how phonetic transforms work, however, the fact that the phonetic distributions for the two categories are not identical to each other in any position indicates that what is happening is *not* that one category is changing into another at the AC-representation level. I view the rough preservation of the variance structure as welcome, but tentative, additional support for the uniformity of this class of process with the allophonic processes discussed in Chapter 3.

### 5.1.1 Empirical predictions

#### 5.1.1.1 One category, one process

We left the framing of the relation between complementary distribution and phonetic rules in Chapter 4 as: “phonetic rules imply complementary distribution, but why should the converse hold?—i.e., that complementary distribution implies or strongly suggests a phonetic rule.” Actually, there are several things about this that need to be cleaned up before we can proceed:

(108) *Complementary distribution* needs to be swapped out for *degree of complementarity of distribution*.

(109) *Suggests* means *increases the posterior probability*.

(110) *Phonetic rule* needs to be swapped out for *magnitude of phonetic transform*.

(111) The last two points imply that *suggests a phonetic rule* needs to be swapped out for *shifts the posterior distribution of the phonetic transform magnitude away from zero*.

(112) Complementarity of distribution can only be evaluated with respect to a particular hypothesized category structure.

The first point is a consequence of the fact that we do not actually perceive categories: we get tokens and infer categories. Thus we cannot ever expect the complementarity of distribution to be any more perfect than our inferred categories. The second point should be self-explanatory by now. The third point is not actually quite right: we *will* be asking a

qualitative, and not a quantitative question; but to start with just consider that the question “is there a phonetic rule” means “how much bigger is the phonetic process than zero,” and the reasoning will be easier. The fifth point is the most important. No structure “just emerges” from any finite data set, apart from the data itself. Complementary distribution is no different. The usual definitions of complementary distribution are based on the simplifying assumption that there is a correct assignment of tokens to a set of phones that could be given by some oracle, and in any particular case we assume that our transcriber has done this assignment right. However, the oracle categories are just one of infinitely many structures that could have been assigned to a finite collection of phonetic observations. Thus, in order to talk about complementary distribution at all, we need to hold an assumption about how the “true” category assignment goes and implicitly assume that such a true assignment exists; it does not simply suffice to be given a phonetic corpus. For the current purposes, I will make life easy and assume that the true oracle phones are a pair of equal variance Gaussians which may differ in location. Each one gives rise to pronunciations in a different context, either  $\bar{x}$  or  $\bar{x}'$ . Simplifying, I hold constant that the learner is attempting to learn the base representation  $\underline{r}$  and the transform representation  $\underline{t}$  for only one single category. Given this, if the locations for the two (true) categories are exactly the same, then I take this to be the case where there really only is one oracle category.

I will call the oracle Gaussian associated with  $\bar{x}$  the category  $A$ , and I will call the one associated with  $\bar{x}'$  the category  $B$ . Complementary distribution is about the probability distribution of contexts in which a particular category appears. Even besides the issues about categories versus tokens, this is a problematic definition if the categories  $A$  and  $B$

are simply defined as those categories associated with the contexts  $\bar{x}$  and  $\bar{x}'$ ! We instead take the context to be the output of some decision rule for the category given a particular observation, under the intuition that complementary distribution asks how well we could use the “apparent” categorization to predict the context. Thus here the context distribution which is to be assessed as complementary or not is the distribution of predicted  $\bar{x}/\bar{x}'$  given the actual category (either  $A$  or  $B$ ).

The degree of complementarity  $d$  is, as in Chapter 3, the symmetrized KL-divergence. Hold constant  $N_A$ , the number of observations that are actually from category  $A$ , and  $N_B$ , the number of observations that are actually from category  $B$ . Let  $N_A^\checkmark$  and  $N_B^\checkmark$  be the number of  $A$  and  $B$  observations correctly classified by the decision rule, respectively, and  $N_A^\times = N_A - N_A^\checkmark$  and  $N_B^\times = N_B - N_B^\checkmark$ . Then we have:

$$\begin{aligned}
 (113) \quad & \left[ \log \left( \frac{N_A^\times}{N_B^\checkmark} \cdot \frac{N_B}{N_A} \right) \cdot \frac{N_A^\times}{N_A} + \log \left( \frac{N_A^\checkmark}{N_B^\times} \cdot \frac{N_B}{N_A} \right) \cdot \frac{N_A^\checkmark}{N_A} \right] \\
 & + \left[ \log \left( \frac{N_B^\checkmark}{N_A^\times} \cdot \frac{N_A}{N_B} \right) \cdot \frac{N_B^\checkmark}{N_B} + \log \left( \frac{N_B^\times}{N_A^\checkmark} \cdot \frac{N_A}{N_B} \right) \cdot \frac{N_B^\times}{N_B} \right] \\
 & = \left[ \frac{N_B^\checkmark}{N_B} - \frac{N_A^\times}{N_A} \right] \cdot \log \left[ \frac{N_A^\checkmark \cdot N_B^\checkmark}{N_A^\times \cdot N_B^\times} \right]
 \end{aligned}$$

Where we are going is to look at the learner’s posterior distribution over transforms  $T_{\bar{x}'}$ , and how it shifts as a function of  $d$ . However, we cannot make any such inferences about what the learner will conclude based on the actual value of  $d$  for any finite corpus: any particular value of  $d$  can be associated with a wide range of posterior distributions over  $T_{\bar{x}'}$ . Instead, we can integrate over all possible data sets for a given pair of categories

$A$  and  $B$  to obtain the expected value of  $d$ . Since a given data set is summarized by the number of correct and incorrect classifications, we have, for fixed  $N_A, N_B$ :

(114)

$$E[d] = \sum_{N_A^\checkmark=0}^{N_A} \sum_{N_B^\checkmark=0}^{N_B} p_{A^\checkmark}^{N_A^\checkmark} (1 - p_{A^\checkmark})^{N_A^\times} \cdot p_{B^\checkmark}^{N_B^\checkmark} (1 - p_{B^\checkmark})^{N_B^\times} \cdot \left[ \frac{N_B^\checkmark}{N_B} - \frac{N_A^\times}{N_A} \right] \cdot \log \left[ \frac{N_A^\checkmark \cdot N_B^\checkmark}{N_A^\times \cdot N_B^\times} \right]$$

The summand is always positive and finite, and, as long as  $p_{A^\checkmark}, p_{B^\checkmark} \geq 0.5$ , then it is increasing in both. The most obvious decision rule is to select the maximum-likelihood category and toss a coin if the probabilities are equal. For a pair of equal-variance Gaussian distributions, the decision bound is a hyperplane at the midpoint between the two means—in our case, the hyperplane normal to  $\underline{t}$  which passes through  $\underline{r} + \frac{1}{2}\underline{t}$ —and  $p_{A^\checkmark} = p_{B^\checkmark} =: p^\checkmark$  is just the total probability beyond the decision bound to the side of the Gaussian which faces the other category. Since this is always the “smaller” side, for  $|\underline{t}| > 0$ ,  $p^\checkmark > 0.5$ . Thus the summand and thus  $E[d]$  increases in  $p^\checkmark$ . There is only one way for  $p^\checkmark$  to increase, however, for fixed variance: increase  $|\underline{t}|$ . Thus if the expected degree of complementarity of distribution increases, then the separation between the categories must have increased.

Now, what does this have to do with the learner? In the limit, as the size of the corpus goes to infinity, the mean values for the two clusters will be  $\underline{r}$  and  $\underline{r} + \underline{t}$ , and it is easy to show that the mean of the posterior on  $\underline{r}^*, \underline{t}^*$  will be Gaussian and centered at a combination of the prior mean with these empirical means (with some small corrections for correlation). Thus, as  $\underline{t}$  moves away from zero, so will the posterior for  $\underline{t}$  shift in that direction. Thus: as complementarity of distribution increases, the magnitude of the



preferred phonetic transform also increases.

This would not exactly be what we were originally after—complementary distribution leads *to a phonetic rule*, not leads *to a larger-magnitude* phonetic rule—were it not for the variable selection parameter  $\gamma$ . Recall from Chapter 3 that this model allows the learner to disable particular phonetic transforms by fixing them at zero ( $\gamma = 0$ ). Recall further that no a priori bias towards one or the other value of  $\gamma$  is necessary. The joint posterior on  $\gamma$  in conjunction with the other hyperparameters it regulates gives rise to a Bayesian Occam’s Razor effect (see Chapter 2). In our case, this takes the expression for the joint posterior from this (ignoring  $\Sigma$ )—

$$(115) \quad f(r, t | X, Y) = \int f(r, t | X, Y, A_0, \Omega) f(A_0) f(\Omega) \, dA_0, \Omega$$

—to this:

$$(116) \quad f(r, t | X, Y) = \frac{1}{2} \int f(r, t | X, Y, A_0, \Omega) f(A_0) f(\Omega) \, dA_0, \Omega \\ + \frac{1}{2} \int \delta(t) \cdot f(r | X, Y, A_{0,\text{red}}, \Omega_{\text{red}}) f(A_{0,\text{red}}) f(\Omega_{\text{red}}) \, dA_{0,\text{red}}, \Omega_{\text{red}}$$

The densities  $f(A_{0,\text{red}})$  and  $f(\Omega_{\text{red}})$  are meant to be the “reduced” densities for the distributions on these hyperparameters wherein the  $\underline{t}$  row of  $A_0$  and the corresponding rows and columns of  $\Omega$  are fixed at zero to yield a degenerate Gaussian. These densities are necessarily numerically greater over the regions where  $\underline{t} = 0$ , because (see Chapter 3) they are obtained as the conditional distributions of the relevant submatrices—satisfying similarity, (see Chapter 2), and thus giving rise to the Bayesian Occam’s Razor. The consequence

is that the posterior probability will be higher for some particular “no transform” solution than for some other solution with a given non-zero  $\underline{t}$  if the following ratio is greater than one:

(117)

$$\frac{f(Y|X, r, 0)}{f(Y|X, r', \underline{t})} \cdot \left[ \frac{\int f(r|A_0, \Omega) f(A_0) f(\Omega) dA_0, \Omega + \int f(r|A_{0, \text{red}}, \Omega_{\text{red}}) f(A_{0, \text{red}}) f(\Omega_{\text{red}}) dA_{0, \text{red}}, \Omega_{\text{red}}}{\int f(r', \underline{t}|A_0, \Omega) f(A_0) f(\Omega) dA_0, \Omega} \right]$$

Even for the most favorable circumstances, the Bayesian Occam’s Razor will give rise to a bias for the simpler solution (via the difference in the determinants of the hyperparameters  $S_0$  and  $\Phi$  for  $A_0$  and  $\Omega$  across the full and reduced cases). In the case where the maximum likelihood solution is close to  $\underline{t} = 0$ , the posterior probability on  $\underline{t}$  (since it is Gaussian) will be reasonably high at  $\underline{t} = 0$ , and the left-hand (likelihood) ratio will be not much less than one. Thus for distributions which are non-complementary because the distributions of phonetic values in the two contexts are exactly overlapping, or close to overlapping, the learner’s bias will clearly favor a system without a phonetic rule (and this is the only possible way the distributions could show low complementarity given the equal-variance Gaussian assumption). This is why complementary distribution leads the learner to posit a phonetic rule, and non-complementary distribution leads the learner not to.

#### 5.1.1.2 Two categories, one process

There is little to be said about the relation between complementary distribution and the choice of the learner to posit two categories or one, except that there is none. What I mean is that, if there are two clusters that the learner is faced with, then the learner may take them to be allophones of the same category if they are in complementary distribution (to

whatever degree). However, the learner may also take them to be two separate categories, regardless. If both clusters are fairly normal and the two are in strongly complementary distribution, then this will require that each of the two separate posited categories have zero transform; nothing more. There is nothing wrong with this solution except that it is subject to countervailing simplicity bias.<sup>2</sup> The tradeoff between the simplicity bias against this solution and the simplicity bias against the one category, one process solution—given that both achieve precisely the same likelihood, assuming the two categories have equal variance—is complex and sensitive to the prior. Thus I will say nothing more about it. However, consider this our base case: the two oracle categories are in perfect complementary distribution, as before.

Now consider the following two cases: (i) category  $A$  systematically lacks any observations from  $\bar{x}'$  but category  $B$  is mixed (we could say that there are three categories, two  $\bar{x}$ , and one  $\bar{x}'$  overlapping perfectly with one of the two  $\bar{x}$ ); (ii) both categories  $A$  and  $B$  are mixed (we could say there are four categories, in the obvious way). Now, clearly, the SKLD will be greater for case (i) than for case (ii). Case (ii) obviously should give rise to two different categories with a simplicity bias favoring no transform, just as in the previous section. For case (i) it is still clear that two categories are needed, but what is more subtle is the fact that positing a non-zero transform that perfectly overlaps an existing category—a kind of “quasi-structure-preservation”—leads to *no* boost in likelihood!

---

<sup>2</sup>The role of BOR in the Dirichlet process mixture model is somewhat more complex than for the variable selection prior. First, there is also a non-BOR simplicity bias effected by the “rich get richer” property. Second, the consideration of the addition of a previously unused category must be made point by point, and so the expression illustrating where the BOR comes from in a DP mixture for a data set of size  $N$  is extremely unwieldy. The idea is the same, though: probability of making use of an already-used category is scaled by the likelihood under that category, whereas for a previously unused category, the probability is scaled by the likelihood *integral* over all possible parameters. If the likelihood is already quite high for the existing category, then it will tend to go down when integrated.

Thus this too should be dispreferred by the BOR.

However, now consider what happens when we tweak case (i) to look like incomplete neutralization by overlaying the additional  $\bar{x}'$  copy of category  $A$  (call it  $A'$ ) not exactly at the same location as category  $B$ , but close to it. Assume that it consists entirely of observations from context  $\bar{x}'$ . As the separation between the two means increases, the best achievable likelihood by a two-category, one-transform model begins to exceed the maximum likelihood solution under a two-category, no-transform model, and there will reach a point, just as in the one-category case above, where the scale tips against the BOR in favor of a phonetic rule. Exactly the same reasoning holds for the case where the covariance differs between  $A/A'$  and  $B$ , and, of course, both ways of being incomplete neutralization—different location and different category shape—will in general be combined. Thus we see that, at a certain point, (which will depend on the particular data set), the increase in likelihood permitted will outstrip the BOR effect induced by the choice to set  $\gamma = 1$ .

All of the foregoing analyses assume that categories nicely hew to the assumptions of the Chapter 3 linear Gaussian model, and of course, this is not exactly true. However, the point should be clear: whatever data makes a “good” incomplete neutralization is different from what makes a “good” two-category, no-transform solution in terms of the likelihood; and, considering only one category, complementary distribution data makes “good” a phonetic-transform solution. Whatever the correct theory of phonetic category maps and  $\oplus$  turns out to be, there may be cases for which there is actually no gain in likelihood, like the quasi-structure-preservation case discussed above. For these cases, we predict that the learner will choose the simpler model if there is one. In general, however,

the learner's preferences will be driven by a combination of fit to the data and bias (including simplicity). This illustration of the corner cases shows how the reasoning goes, and suggests, if allophony really is phonetic and incomplete neutralization really is inferred by the learner in the face of just the types of phonetic data illustrated above, that the current phonetic category conception shares at least some important properties with the correct one. Stepping back further, it gives a clear picture of why and when allophony (and incomplete neutralization) should be a late phonetic rule: under the current model, these solutions lead the learner to a better tradeoff between faithfulness to the data and markedness of grammar.

#### 5.1.1.3 Two categories, one categorical process

We have showed that, when two clusters both seem to contain observations from both of two contextual environments, the best result would be two separate categories, and zero contextual transforms. This is one case of non-distribution that would therefore lead the learner *not* to posit a transform-. We have also showed that cases where the two clusters are in totally complementary distribution are predicted under a single-category, single transform solution—that many cases, in particular those where the two are plausibly Gaussian with the same variance, would lead to single-category, single-transform systems. Thus, complementary distribution, all things being equal, leads the learner to non-structure-preserving late phonetic rules. Finally, we also showed that there were exotic “quasi-structure-preserving” solutions that could be supported in case a single cluster was largely devoid of observations from some particular environment, but that these would

be dispreferred by the prior. We showed what the intermediate cases looked like: two categories and a single non-zero transform would be preferred in a whole host of cases in between classical strict allophony and quasi-structure-preservation, all of which we might call “incomplete neutralization.”

Now we would like to say something about what the grammar *would* do in the face of the quasi-structure-preservation data. One option is nothing: as we said, it is just a case where one region happens to be missing data from one particular context; there is nothing special about this, and so nothing more really needs to be said. It is usually thought, however, that phonological grammars nevertheless *do* say something about this. Without an explicit prior for phonological grammars, it is difficult to say much more. Furthermore, as was discussed extensively in the preceding chapters, there are two classes of solutions to this problem, depending on the theory and depending on the circumstance: simply state a static restriction, to the effect that “this category never appears in this context,” and let that be the grammatical knowledge; or implement a causal explanation, to the effect that “this category becomes this other category in this context.” Consider these two types of statements and the predictions they make:

1. We subsequently consider sequences in which a segment is very likely some category  $A$  and its surrounding segments are very likely the impossible/neutralizing context  $\bar{x}$  to “not fit” very well (both cases)
2. We subsequently consider the neutralization target  $B$  to be one of two lexical segments,  $A$  or  $B$  (second case)

Clearly, in both cases, the grammar is more restrictive with respect to the set of AC-

sequences that will be tolerated. This is a good thing, because any licit AC-sequence will have higher likelihood (see Chapter 2 for the short discussion of restrictiveness in probabilistic inference). If the grammar is variable-length, this will presumably trade off against simplicity, which will disprefer the additional constraints or rules needed to guarantee this restrictiveness (again, see Chapter 2). However, one thing is clear: so long as it is possible to do so, the learner will have a motivation to ban segment  $A$  in context  $\bar{x}$ . This has nothing to do with morphological pressures, which are often intuitively thought to be the principal motivation for the grammar “explaining” the absence of a particular phone sequence. See Jarosz 2006 for models making this point in another way.

I will not attempt to adjudicate between these two types of formulations; as discussed above, using an Optimality-Theoretic grammar, it is possible for a surface constraint to give rise to the same ambiguity about the underlying identity of a segment in the context  $\bar{x}$ , without additional (more complex) grammatical apparatus, if other morphological pressures demand it. Similarly, under a solution which forces all instances of  $A$  in context  $\bar{x}$  to be realized as  $B$ , there is no need for an explicit surface constraint banning  $A$  in that context (except in the case, irrelevant here, where other processes create new instances of  $A$  in that context).

It is worth pointing out, however, that, the preference for restrictiveness notwithstanding, morphological pressures can also be relevant. In particular, the learner may have reason to collapse two similar strings as instances of the same morpheme, with a common underlying representation (see Chapter 2, Chapter 4). In this case, restrictiveness will encourage the formulation of grammatical constraints that rule out free variation in the realization of this morpheme, and restrict each of the different variants to the contexts in

which they occurred. The effect of this is that, for a given instance of the neutralization target  $B$  in the context  $\bar{x}$ , other facts about the surrounding environment will boost the posterior probability of either (i) underlying representation  $B$ ; or (ii) underlying representation  $A$  and a grammar requiring, not merely tolerating,  $B$  in the context  $\bar{x}$ . Information of type (ii) would boost the probability of either a quasi-structure-preservation solution or a phonological solution; we would rely on the variable selection prior more strongly dispreferring the quasi-structure-preservation solution than the phonological grammar prior disprefers the neutralization solution, if we wanted to ensure we get the “classical” phonological neutralization solution back. The architecture of grammar then becomes very important: cyclic reapplication following the SCC (see Chapter 4) and process interactions (see below), along with the homoscedasticity assumption of our linear Gaussian model, or whatever is actually to be said about the phonetic interpretation of phonetic transforms, become crucial pieces of evidence pushing the learner towards one solution or the other.

### 5.1.2 Summary

Incomplete neutralization is not a mysterious phenomenon, nor, I have argued, should it ever have been. It is simply allophony that never quite got its ducks in a row; but the phonetic grammar is forgiving. As far as the grammar is concerned, it is simply allophony. The only sticking point arises when we consider the fact that the “quasi-structure-preservation” solution coexists with the phonological solution in the set of possible grammars. The reason that late rules are not structure-preserving, according to this theory, is entirely due to a bias for simplicity. I have called the previous section “Empirical predic-



tions” not because I believe that we could test any specific predictions tomorrow; absent a particular prior for phonological grammars, what exactly the system learners will arrive at remains something of a mystery. Rather, I believe simply believe that I have laid out the corner cases well enough that, with a more complete theory, we *could* make predictions, and we could start to constrain such a theory by examining changes currently going on for which we have phonetic corpora. The German incomplete neutralization facts appear to be stable; unless some other process comes into the language with which it could potentially interact, there is no reason to think that learners are going to upend its incompleteness any time soon. However, if the two categories were to change in all positions to become more equal in shape, then, under the current instantiation of the theory, the learner would have much more trouble telling the incompletely neutralized allophones apart. Learners would eventually be predicted to make a saltatory leap from a phonetic to a categorical process. This is presumably an unmarked case for the phonemicization of an allophone and the introduction of a categorical process into a language.

## 5.2 Phonetic process interactions

### 5.2.1 Predictions

We come to the last topic to be explored in this dissertation: interactions among phonetic transforms. *Interactions*, broadly construed, are simply those considerations that need to be taken when two or more mappings are thought to be potentially applicable to a particular input—and yet there is only one output. Do the mappings compose, with the output to one being the input to the next? If so, which way do they compose (or does it not

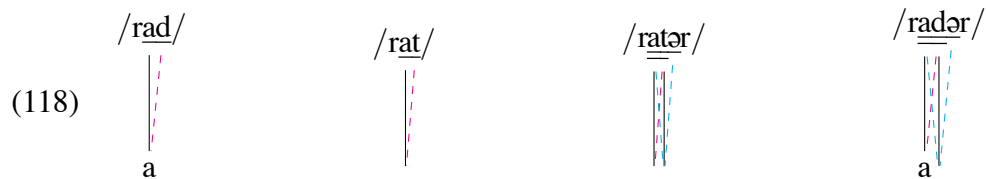
matter)? Is there a different kind of combination at play, whereby one input is submitted to several mappings, and then some other function, or set of functions, is used to combine the outputs? The issue turns only tangentially on whether a grammar is represented as the specification of a set of grammatical sub-mappings, or as a set of output constraints, or whatever else we might think. Any useful large collection of mappings will inevitably be constructed out of combinations of elementary mappings, and phonological grammars, broadly construed, are no exception. That is because by “represented” we generally refer only crucially to some structure followed by the learner; it makes no sense to think about the receptive or productive computations taking place by “computing” output constraints without “computing” outputs!—and it is hard to imagine either of these computations, given their wide range of possible inputs, not somehow being structured into a combination of elementary processes (deletion, spreading, footing, elementary transducers, or whatever else we might propose). Rules about how these elementary operations combine are always necessary, whether to the net operation of a grammar learned as individually specified grammatical processes or to the theory of Gen.<sup>3</sup>

Consider, for example, the Canadian Raising pattern discussed several times now: *write, ride, writer, rider*: [rjt], [rajɖ], [rjər], [rajər]. There are essentially two processes involved: a process changing the vowel quality (Raising) and a process shortening the closure duration and roughly neutralizing the voicing contrast for the following coronal

---

<sup>3</sup>This cuts both ways: for phonetic grammars, although we have assumed that the prior acts as if individual processes are directly specified, we could equally imagine a theory where phonetic transforms were specified indirectly via constraints. The questions about how elementary phonetic transforms combine to give rise to the net result would not change; what would potentially change is how any given grammatical mapping is to be decomposed (and that means that the implications of any particular empirical case might change). In addressing particular cases below, I will continue to assume that two general phonetic processes that can be identified independently are objects that will need to be combined using  $\oplus$ .

stop (Flapping). Assuming the diphthong is to be taken as a single segment, Raising is clearly non-structure-preserving, and thus obviously phonetic under the current theory, (where “obviously” stands in for all the qualifications outlined above). The Flapping rule is phonetic for similar reasons.<sup>4</sup> The interaction between these two processes is exactly as predicted under the strong phonetic transform hypothesis, whereby only categorical information can feed phonetic processes (—in the same EOD block, if the rules are level ordered).



Under the current hypothesis, this kind of interaction is completely prototypical. The input to the phonetic computation is of a different type than the output, thus composition is impossible, and the only choice for application is *simultaneous application*, followed by reconstruction of the output using another set of functions (in this case, the sequential relation is the same, but implicitly there is some kind of phonetic sequencing function that combines the outputs; and in general, for transforms applying to the same segment,

<sup>4</sup>If Flapping is a phonetic transform, we predict that the neutralization is highly likely to be incomplete. The phonetic facts are nuanced. The preceding vowel duration, which is a near-universal correlate of voicing, is incompletely neutralized. The closure duration differences are miniscule and not consistently in the same direction across experiments, and thus there is very likely no difference at all (/d/ – /t/ –4 ms: Charles-Luce, Dressler & Ragonese 1999; +0.9 ms: Herd, Jongman & Sereno 2010; +0.4 ms: Braver 2011; Braver also finds only a 0.7% difference in percentage of voicing into closure; neither of the latter two studies report variances, but the histogram given by Herd et al. seems to suggest close alignment of distributional shape). The qualitative difference between the two dimensions may be due to a floor effect in the closure duration (stop closure durations do not generally go below 25 ms, and the closure durations in these sets of results are close to this minimum), but this leads to a potential alternate explanation: the process is truly neutralizing, but the vowel shortening is phonetic and allophonic. This explanation is particularly salient in light of the fact that the much shorter duration of [j] as opposed to [a] is usually lumped in with the effects of Raising. However, suppose this is correct: Shortening must still be sensitive to the underlying voicing status of the flap, and, since Shortening is phonetic, under our theory, Flapping still *must* be phonetic (or at least must follow the relevant phonetic EOD block).

we also use  $\oplus$ ). This is different from the previous approaches to Canadian Raising and related phenomena. The surface-constraint approach struggles because it is different: in particular, it struggles because the environment for Raising cannot be correctly stated with respect to the outputs. The derivational approach is fine, but it works differently: the environment for (crucially) Flapping is stated over an input that is the output of Raising. This is not necessary, but it is forced under the derivational theory. Instead, the current approach states that the input for both rules is the same object, the output of the categorical phonology, which I have called the AC-representation.

Chomsky & Halle 1968 argued against earlier attempts to introduce simultaneous application. One reason was that it could not handle all patterns: it undergenerated.<sup>5</sup> Another reason was that, although most cases of simultaneous application can be given an analysis under a derivational theory, the one case of a simultaneous application pattern that is impossible for derivational theories does not seem to be attested: thus simultaneous application never really needed and it seemed to overgenerate. Putting these two things together, simultaneous application was neither necessary nor sufficient to handle all of the

---

<sup>5</sup>Most of what they discussed there and elsewhere in this regard falls outside the scope of phonetic rules. However, in one example, they do invoke Raising. The interaction is with Pig Latin, however, which is a language game and thus makes the case somewhat more difficult to reason about. The example goes like this: all English speakers with Raising who know Pig Latin will pronounce *sight* → *ightsay* with [j]; but some speakers will pronounce *sigh* → *ighsay* with [aj], while others will pronounce it with [j]. Simultaneous application of Raising and Pig Latin would predict only [aj], never [j], since the voiceless [s] does not follow [aj] underlyingly. It is hard to know what type of rule Pig Latin is, but the possibility of [j] would imply that some speakers treat it as reordering the segments in a way that phonetic interpretation of the individual segments can see. Level ordering might help to explain this difference, with [aj] being amenable to a “Late Pig Latin” solution, if Bermúdez-Otero is correct that Raising does not apply at the word level, (see Chapter 4), or if level ordering could give rise to a similar effect because of failure to resyllabify. Early Pig Latin would be an application of the same rule at a narrower HOCd. The Pig Latin rule would in any case need to be able to rearrange segments and resyllabify on the output of phonetic interpretation. The “real” example they give, following Joos, of “Dialect B,” in which Flapping precedes Raising to yield [tjprajər] has been argued never to have existed (Kaye 1990) on the simple grounds that Joos claims the speakers were schoolchildren in 1942, but Kaye could find no trace of them as adults in 1990. If these speakers did ever exist, though, for them, Flapping must in fact have been categorical Voicing, [tjprajdər].

phonological mappings. With our newly restricted focus on allophonic (/incompletely-neutralizing/quasi-structure-preserving but late) rules, I will reexamine the question of sufficiency in this question.

The question of necessity is interesting as well, but I have less to say about it. The case given by Chomsky and Halle that they said would make it necessary was like this: given rules of the form  $A \rightarrow Y/-X$ ,  $B \rightarrow X/-Y$ , in a mutual feeding relation, the prediction of simultaneous application is that  $ABY$  sequences will surface as  $AXY$ , and  $BAX$  sequences will surface as  $BYX$ . The two rules are both triggered by the underlying, not the derived, environments, and so applying one rule has no effect on the operation of the other. This pattern is impossible under either ordering if the combination is by direct composition: either the output  $Y$  of the first rule should trigger the second, or the output  $X$  of the second rule should trigger the first. However, the resultant pattern, they said, interestingly does not exist. Under the current theory, a clearer way to write these two hypothetical rules would be as  $A \rightarrow y/-X$ ,  $B \rightarrow x/-Y$ , where  $y$  and  $x$  are two phonetic representations not interpretable as  $Y$  or  $X$ . Such circumstances are rare to begin with, and I can find no relevant examples; however, the current theory predicts that this is not a systematic gap.

### 5.2.2 Possible counterexamples

Rubach 1984, Booij & Rubach 1987 propose that the end-of-domain rules applying at the word level, the *post-cyclic rules*, are distinct from those applying at the end of the derivation in toto, including across word boundaries, the *post-lexical rules* (see Chapter 4). We have already seen in Chapter 4 that in a model with cyclic spellout, any kind of

interactions between end-of-domain rules are problematic for the theory that these are all phonetic interpretation rules; but Booij and Rubach's post-cyclic rules are extra problematic: not only can they interact with post-lexical allophonic rules, they can interact with each other, they can tolerate exceptions, and they can be neutralizing. What separates them from the cyclic rules is that they can apply morpheme-internally, and they do not feed any of the cyclic rules in their own domain. There problems are clearly real: the Surface Palatalization rule adding noncontrastive palatalization to consonants preceding [i], but the post-cyclic rule of Retraction taking [i] to [ɨ] can bleed it. This rule has exceptions and also itself crucially follows the *r*-Spellout and Yer Deletion post-cyclic rules. A reanalysis of Polish phonology which avoids reference to the problematic post-cyclic rules would be going too far afield, since there are already questions about interactions between end-of-domain rules; we must accept for the moment that there may be indeed an end-of-domain block which is not phonetic, limiting the scope of the theory substantially.

However, they do report an interaction between two post-lexical Polish voicing rules. Polish has a rule of Final Devoicing which applies word-finally to all obstruents, as in *sad*, "orchard," with a devoiced final consonant, versus the nominative plural *sad+y*, where the *d* is voiced. Polish also has a rule of Regressive Voicing Assimilation by which obstruents assimilate in voicing to following stops or voiceless fricatives, morpheme-internally and across morpheme boundaries, as in *proś+b+a*, "request" (nominalization), with *ś* pronounced as voiced due to the following *b*, and to all following obstruents across word boundaries, as in *kryzys gospodarczy*, "economic crisis," with voicing on *s* due to the following *g* (we will discuss the value of collapsing the two cases together in a moment). Now, what happens if a word-final voiced segment, which ought to be subject to Final De-

voicing, is also followed by a voiced segment, which ought to trigger Regressive Voicing Assimilation to voice a voiceless segment? The result is that the segment is voiced, as in *sad wiśniowy*, “cherry orchard,” with the initial [v] in *wiśniowy* evidently responsible for the fact that *d* is voiced.

For these two rules to be phonetic is not a problem for the current theory: the environments “word-final” and “following voiced obstruent” should sum using  $\boxplus$  and the two transforms should thus combine using  $\oplus$ . The segment is predicted to be both devoiced and voiced, and would probably be somewhat distinct in voicing from underlying voiced stops in other positions. However, this is somewhat strange, as it requires that underlying voiced stops be subject to assimilation to a following voiced stop. This is not implausible, but presumably the phonetic difference would be rather small normally. On the other hand, at least if the transcription is to be trusted, then re-voicing is evidently sufficient to push a largely voiceless segment to being perceived as voiced. It is clear that we predict that these re-voiced segments ought to (almost surely) have some different voicing status from both devoiced and underlyingly voiced segments, but we really do not want to predict that the difference between re-voiced and devoiced segments is barely detectable, which is what would be predicted under a naive view of  $\oplus$ .

Now, we do not know exactly what the  $\oplus$  operator is, and so it might be possible that phonetic voicing would combine with phonetic devoicing in a way that would yield the desired effect, which is (again presumably) that re-voiced segments are more like voiced than devoiced segments. One approach would be to say that  $\oplus$  is not generally commutative, justifiable if here  $\boxplus$  is not commutative, perhaps because the context inside the word domain (“word-final”) must be “added” first. Another approach would be to recall from

Chapter 3 that  $\oplus$  appears to do something more than just add raw acoustic measurements, (naturally), but in particular that it does something consistent with some kind of rescaling of the bounds of the phonetic space (see above also). Even that appears to be problematic, however. Recall that in Chapter 3 we showed data from Kalaallisut suggesting that Post-Coronal Fronting and Uvular Retraction, when **both** applied, had effects on vowels that were lawfully related to their effects when only **one** applied. However, in that case, we saw that the effect of either was **smaller** acoustically when they combined, and that led us to a particular conjecture about transforms as scaling wherein the number of transforms applying would predict the degree to which the phonetic space is scaled down. Here, on the other hand, the effect of Regressive Voicing Assimilation to combine in the appropriate way with devoicing, the phonetic space needs to be scaled *up* under the application, as versus the non-application, of the transform.

Now consider that there is yet a third voicing assimilation process, Progressive Voicing Assimilation, which applies to sequences of two obstruents within morphemes and across morpheme boundaries, (but not across word boundaries), when the first is voiceless and the second one is a voiced fricative, as in *bitw+a*, “battle” (nom. sg.), [bʲitfa], versus *bitew+n+y*, “warlike,” [bʲitevn]. Notice the complementarity of environments word-internally with respect to Regressive Voicing Assimilation: that rule has a special exception when the second consonant is a voiced fricative, which is exactly the case where Progressive Voicing Assimilation applies. The way that Booij and Rubach handle this is by ordering Progressive Voicing Assimilation before Regressive Voicing Assimilation and letting the one bleed the other (if Regressive Voicing Assimilation applied first then we would get \*[bʲidva]). In fact, the rules *must* be ordered like this because Progressive



Voicing Assimilation is a post-cyclic rule and Regressive Voicing Assimilation is a post-lexical rule. We cannot do this here if both rules are phonetic. Furthermore, if both rules apply, then we expect something more like \*[bʲidfa]. One solution is simply to keep the environment for Regressive Voicing Assimilation as we state it above, explicitly excluding the fricative case.

Another approach would say that voicing assimilation is simply qualitatively different from the other phonetic rules we have been looking at so far. Following the intuition that Regressive Voicing Assimilation's yielding of either voicing or devoicing, depending on the trigger, suggests a kind of spreading, we might posit a simple deletion of the target voicing feature and then some interpolation of voicing according to the remaining marked feature, in a manner following Cohn 1990. We could then retain the simultaneous application of Progressive and Regressive Voicing Assimilation: both voicing features delete word-internally when the second is a voiced fricative, and the result is voiceless not because the first was voiceless but because voicelessness is the default; this is somewhat strange, however, since, intervocally, voicing ought to be the natural interpolation. If it is correct, though, then we still need to say something about the case where a gradient process of Final Devoicing could also apply: does the voicing information on the word-final obstruent get removed by Assimilation, filled in phonetically and then modified in a gradient way using  $\oplus$  by Devoicing, or does Devoicing fail to apply because the feature has been removed? According to the transcription, the result is consistent with the latter, but if these processes really are gradient then the question is worthy of further investigation. Generally speaking, the introduction of a second type of late/phonetic operation, feature removal, demands that we say how it interacts with itself and with  $\oplus$ , in particular

cases or in general.

In sum, the logic of the problem is as follows: Regressive Assimilation arguably, and certainly according to the standard description, counterfeeds Final Devoicing, which *is* the type of rule interaction predicted for phonetic processes applying to different segments (see above), but which is **not** what is predicted when the processes apply to the same segment, as in this case. Rather, we predict something which does not have any natural correspondent in the classification of categorical rule interactions: combination with  $\oplus$ . If it is actually not true that RA genuinely negates FD, then we are faced again with the question of exactly what it means for the two to combine with  $\oplus$  in the usual way. As we have some doubts about the validity of this phonetically, at least given the coarse description of the output, we raise the possibility that RA is not a phonetic rule combining with  $\oplus$ . It would not do to say that RA is a phonological rule that makes the resulting offending non-devoiced/re-voiced segment [+voice], because this would do nothing to solve the problem. The interaction with Progressive Assimilation is arguably less important, because the only issue, if the two are both phonetic processes combining with  $\oplus$ , is that the prior distribution on complex environments for phonetic rules might be thought to disprefer the statement of the RA environment which explicitly excludes the fricative case on grounds of simplicity; still, this is a possible grammar for which the available data in favor ought to be abundant, so it seems likely that the problem would dissolve.

Before making any commitments, let us consider the facts from Dutch, which are almost exactly the same (Zonneveld 1983, Grijzenhout & Krämer 2000). Dutch also has a rule of Final Devoicing (*pad*+*en*, “toad (pl.),” [pdən], versus *pad*, “toad,” pronounced similarly to [pt], but seemingly a phonetic process: Ernestus, Lahey, Verhees & Baayen 2006).

The rule does not apply only at word boundaries, but also finally in narrower phonological domains, as in *goud+achtig*, “gold-ish,” [xut.x.təx], the outer domain of which is evidently the main stress domain, but (at least according to Grijzenhout & Krämer 2000) does not trigger internal resyllabification, making the internal domain evidently a stronger one than the narrowest domains of affixation, (for which Grijzenhout reports resyllabification), but weaker than the word-level domain. Dutch also has Regressive Voicing Assimilation applying across the board to assimilate obstruents to the voicing on following stops, as in *eet+baar*, “edible,” given as [edbar]. Furthermore, Dutch also has Progressive Voicing Assimilation, which applies, as in Polish, to voiced fricatives after voiceless obstruents. Unlike in Polish, both of these processes apply (as stated) across word boundaries. According to Grijzenhout, RA does not apply in the case of the second consonant being a voiceless fricative, unlike in Polish; instead, PA is extended to this case too. This would predict, for an underlying voiced–voiceless fricative sequence, that we would get voiced–voiced; RA would predict voiceless–voiceless. Surprisingly, Grijzenhout reports voiceless–voiceless in this case (*vriend+schap*, “friendship,” with the *d* voiceless). However, since this is a sufficiently strong boundary to be an FD environment, she attributes this to FD applying *first* to devoice the *d*; then PA applies. Examples from weaker boundaries or morpheme-internal cases could help us sort this out, but the L(I) suffixes in Dutch are all vowel-initial, according to Booij 1977, and morpheme-internally, the fact is that on the surface there are no voicing mismatches, and even if there were underlyingly, there would be no way to tell which segment was voiced and which voiceless. I will return to this in a moment.

For now, I put aside the case where the second segment is a voiceless fricative, and

just consider the remaining cases. The interaction between Dutch RA and FD is just like in Polish. If the sequence is underlyingly voiced–voiced, and the environment is both an RA and an FD environment, then the result is voiced–voiced—either by blocking or re-voicing—whatever the phonetic facts turn out to be: the compound *leef+baar*, “liveable,” is reported as [levbar] (in spite of the quirky spelling as *f*, which reflects neither the voicing on the underlying source morpheme, *lev-*, (as in *lev+en*, “to live”), nor the surface pronunciation).

The interaction between Dutch PA and FD is now relevant; in Polish, there is no possibility for such an interaction. The relevant case is underlying voiced–voiced fricative. The relevant example is *raad+zaam*, “advisable,” which is said to surface as [ratsam]. The implication is that FD does indeed feed PA, consistent with the analysis of the voiced–voiceless fricative case. This is a problem. The logic is much like the cases reviewed in Chapter 4: there we reviewed the problem of interpreting higher-order cyclic domains as phonetic interpretation domains, which was a problem because the categorical information sometimes needed to be preserved. We said that it could in fact be preserved in those cases, by a weakening of the theory to allow recovery of categorical information from gradient information when it at least exists; but that after all the cyclic domain did not need to be interpreted as a phonetic interpretation domain. Now we seem to be faced with a much more serious problem. There is no issue about cyclic domains to be deferred, and we are facing the case where the reason we take the categorical information to be absent is because we suppose that it was never there in the first place. After FD, the only categorical information that is recoverable should be the underlying +voice value, but now we need the voicelessness to spread to the assimilated segment.

Now consider again the three types of solutions laid out for the Polish problem. First, that RA is typed just like FD: the two processes combine with  $\oplus$ , although we then need to qualify exactly what the effect of  $\oplus$  is. This simply denies that RA fully negates the application of FD. Second, RA is not like FD: the two combine in some other way, and in particular, RA is a deletion process, and deletion processes *are* allowed to bleed phonetic transforms. A third way which would *not* be applicable for RA would be to give up: to say that RA is a phonetic process that combines with  $\oplus$ , but it can block the application of FD, in this case not because a phonetic processes can supposedly be sensitive to gradient information in the contextual environment, but in fact because they can apply or fail to apply on the basis of precise phonetic value output by another process of the phonetic grammar. This would not work for the same reason that making RA phonological would not work: it makes the wrong predictions.

Now apply this logic here. The problem is not that PA needs to undo FD, but rather that FD seems to need to precede PA. The other difference is that there are two different targets involved. The analog of the first approach would be to say that, by using  $\oplus$  to combine PA voicing *the second segment* and FD *on the first segment*, we somehow wind up with a voiceless *second segment*. This does not have much sense to it. There is not even the slightest independent motivation for making FD into a qualitatively different type of process. As for lifting some limitations on  $\oplus$  process interactions, this will do here: if PA is an  $\oplus$  process, then a simple admission of gradient environments into the theory, in violation of our research strategy, would say that, in the voiced–voiced fricative case, FD first applies to give a devoiced segment, and then PA takes as input a context equal to, not the value of the voicing feature, but some gradient phonetic value; it would perhaps adjust

the voicing in proportion to the voicing on the preceding segment.

We could get something like this make more sense if, again, PA is qualitatively different, and not an  $\oplus$  process. The way this worked for RA was that RA could bleed FD if RA **removed the voicing feature on the segment that is the target of FD**. If PA removes the voicing feature on the second segment, however, then that segment is crucially *not* the target of the transform FD. The result follows if PA removes the voicing feature, then FD applies to affect the voicing on the first segment *before* interpolation takes place. This is “before” in the informational sense: the output of interpolation and FD together is the predicted output of interpolation applying to the output of FD; FD applying to the output of interpolation would give voicing on the second segment but voicelessness on the first segment. The phonetic facts may actually turn out to be subtle, but, if they are just as reported, then, to make this “before” even more natural, we might say that it actually could not have been any other way: the second possible computation, where FD applies to only the first segment, is illicit. This would follow if the segmental boundary were no longer applicable in the interpolation case—FD treats the voicing information across the two segments “as a unit”—and so there would actually be only one way to compose the two. This, in turn, would follow under an autosegmental phonetic model where the phonetic voicing information did not need to be associated uniquely with one segment, but could be linked to two segments simultaneously, and if in fact in this case it had to be for whatever reason. This is largely consistent with what we said about RA, but there is one issue: there we said that RA was deletion of one of the two segments’ voicing features, but FD was blocked because the voicing feature was gone. Here we are saying that PA is deletion of one of the two segments’ voicing features, but FD is not

blocked because the voicing feature is not gone. Evidently, if it is the case that FD takes the segmental boundary to be inapplicable for the purposes of **what gradient information it applies to**, it is *not* the case that FD takes the underlying segmental division between the first-segment voicing feature to be inapplicable for the purposes of determining **whether it applies**. Further manipulation of our intuitions to make this state of affairs appear to follow from something is not necessary—we can simply leave this to be the fact of the matter in this case. However, it may be worth investigating the possibility that this is *only* the fact of the matter in this *particular* case because the environment for FD crucially refers to the boundary following *the particular segment in question*. There are other ways to look at it that might make it a general fact; but we are too deep in speculation at this point.

To sum up: the Dutch and Polish cases both show interactions between what we would like to say are phonetic processes (in the case of FD, for empirical reasons). However, they show two problematic patterns for the present theory: RA either shows a potentially problematic effect of  $\oplus$  in both languages, virtually undoing the effect of FD on underlying voiced segments, in spite of the fact that there is no good reason to say it applies to voiced segments in any way, much less in this way; and PA in Dutch seems to be fed by FD. These two patterns together suggest a different approach to assimilation processes, one in which, although they may be late, they are *not* phonetic transforms that combine with  $\oplus$ . Regressive Voicing Assimilation in Russian stands in the same relation with respect to Final Devoicing as these two cases of RA, and nothing more needs to be said, although some of the details are different.<sup>6</sup>

---

<sup>6</sup>First, RA in Russian spreads until it reaches a vowel; sonorants are optionally either transparent or

I also mentioned a few cases put forward by Booij and Rubach, and, in particular, the problematic interaction between Retraction and Surface Palatalization in Polish: despite the apparent lateness of that Retraction rule, the solution that would be forced upon us here would be that Retraction is a rule of the categorical phonology. Although we have said nothing about lexical exceptions in the present theory, if we did bar phonetic rules from tolerating lexical exceptions, such a solution would be further forced upon us, because Retraction has lexical exceptions—but of course there is really no way to force an outright ban on effective lexical exceptions to particular rule environments as long as the phonetic grammar can perform outright deletion, because there is always the possibility that the AC-representation contains a segment which does not surface and only serves to

---

affected. Second, [v] behaves anomalously, only triggering RA where FD has applied to it. RA and FD, however, affect it obligatorily. Most analyses treat [v] as being in some sense more sonorant-like underlyingly, but, as discussed at length in Kiparsky 1985, the puzzle is how to trade out its non-obstruency (no obligatory FD/RA) for obstruency (obligatory FD/RA, implementation as a fricative) in a way that avoids the unwanted effect of having it trigger RA. The solution of Hall 2007, following Avery 1996 and Rice 1993, is to posit that voicing on obstruents is simply not the same as voicing on sonorants. Russian [v] is anomalous in bearing neither the sonorant (SV) nor the obstruent (Laryngeal) voicing feature, and so it does not generally spread its Laryngeal feature. It is endowed with such a feature either late—prior to RA—in the case of FD/RA, which give it a Laryngeal feature, or very late—after RA—otherwise. The problem if all of this is simultaneous is that cashing out [v] as an obstruent for the purposes of FD will not be allowed to occur before it is cashed out as an obstruent generally. A solution similar to Kiparsky 1985 and Hayes 1984 whereby [v] is a sonorant (i.e., marked with SV), and SV-bearing segments do undergo RA and FD (here, processes that affect the Laryngeal feature only). I differ from these accounts, which treat voicelessness uniformly as the result of having a bare Laryngeal node, in, of course, asserting that it is the filling in of phonetic detail for Laryngeal that gives rise to both RA and FD. This detail may be optionally interpreted on SV-marked segments. The interpretation of [v] as an obstruent involves removing the “prophylactic” SV feature, which may be done simultaneously with RA and FD, and then forces the interpretation of the Laryngeal information. All this is more about the representation and phonetic implementation of [v]; the rule interactions are not a problem in any case. The question about Russian RA that bore on the issue of whether the structuralist phoneme should be maintained, namely, whether it should be split into two different rules depending on the phonemic status of the output, is answered here in the negative. This says nothing about whether the structuralist phoneme *is* in a sense maintained, by which I mean whether or not the result of voicing via RA is actually allophonically distinct from the corresponding underlyingly voiced segment, where those exist. There *would* be a reason to assert this if RA were due to a phonetic transform, which would predict that the result of RA would only be like the underlyingly voiced segment by an unlikely accident; but, although above I suggested an interpolation view of phonetic spreading, it is also possible that spreading does not necessarily lead to non-structure-preserving gradient outputs, but will rather preserve the interpretation of the Laryngeal feature idiosyncratic to the given segment if it has one. Given that transforms like FD can affect gradient detail on several segments at once, there seems to be some tension here, but I defer the issue to a worked out theory of phonetic spreading and phonetic features.



block or trigger the application of a phonetic rule. (The analysis of synchronic Yer Deletion as floating segments by Rubach & Booij 1990 is a proposal that could be potentially reconstrued along these lines.)

One more potential case of interaction of allophonic processes, brought to my attention by Andrew Nevins, is the supposed Laurentian French ATR harmony pattern described by Poliquin 2006. It is an accepted fact that LF has an allophonic Laxing alternation in closed syllables (*petit–petite*, [pt<sup>s</sup>i]–[pt<sup>s</sup>t]), except before voiced fricatives (*église*, [eliz]). Poliquin claims that LF also has a Regressive ATR Harmony rule which, depending on the speaker, can give rise to various exotic patterns such as [smilit<sup>s</sup>d] and [similt<sup>s</sup>d], in addition to across the board harmony, as in [smlt<sup>s</sup>d], for *similitude*. The exotic application facts are almost universally disputed by native speakers, (M. Gagnon, M. Brunelle, p.c.), and Poliquin’s phonetic studies are done over extremely small samples, and so further study is needed. However, if we take at least the existence of some kind of Harmony process at face value, whatever vowels it happens to be able to affect, it is a problem if the process of harmony is seen as taking the output of Laxing as input. The obvious solution is to give exactly the answer as for the RA/PA case: spreading is different. Furthermore, since Laxing applies at the source of the ATR feature, the (only) way to combine it with a sharing of the ATR feature is to apply Laxing to all vowels that share the ATR feature. This forces us into saying that there actually is a phonetic ATR feature (by which I simply mean “dimension”) that gets modified independently of the place information. This is just as in the Dutch and Polish cases: we need to say that there is a mechanism by which the ATR information is shared across vowels. However, this ATR feature does not need to be present, but then subsequently deleted, on the preceding vowels in order to state Harmony,

as one might think following the analyses above (see Chapter 3 on contrastive specification to see the motivation for asking about this). What is crucial is merely the linking of the ATR feature to all the preceding vowels.

Finally, although I have repeatedly deferred the issue of late, “phonetic” deletion above (in Chapter 4), it comes up in an allophonic interaction case in Catalan (Mascaró 1976). Catalan Regressive Nasal Place Assimilation is counterbled by Cluster Simplification, which deletes word-final elements of homorganic clusters, in just the way that would be expected if both are phonetic: *venk*, “I sell,” is [beŋ]. The restriction of CS to homorganic clusters is to exclude cases like *lp*, but this restriction must be lifted for nasals in order for the simultaneous application to work. If deletion is like assimilation in being qualitatively different from phonetic transforms, (as suggested by the analysis of assimilation as involving delinking), then this case fills in a possibility we have not yet seen: the trigger of the NPA transform, rather than the target, (as in the RA–FD case), is removed (in the PA–FD case neither trigger nor target of FD is deleted). In that case, FD failed to apply. Here NPA does apply. The additional issue raised by this case is that NPA can apply across word boundaries and is fed by CS, as in *venc vint pans*, “I sell twenty loaves of bread,” [bɛŋbimpans]. This demands a level ordering solution wherein NPA applies separately at two HOCDs (see Chapter 4), consistent with our discussion in Chapter 4 of these as being the only cases where allophonic feeding should be allowed. It also suggests that not even a trace of the adjacent intervening word-final simplified segment remains after deletion, (here the *t* in *vint*), intuitively consistent with the idea that deletion is qualitatively different from gradient transforms.

### 5.3 Statistics in linguistics

To sum up the chapter: the association between true structure preservation and the phonetic component is explained by an exegesis of the learner's preferences, with incomplete neutralization folded into this category. No complete evaluation measure is available yet for the phonological grammar, and thus for true neutralization. Quasi-structure-preservation is possible but will never emerge. There should be no feeding, (in the sense of feeding or bleeding), among phonetic processes in the same EOD block. As far as I can determine this is empirically confirmed, granting that spreading, and by extension presumably deletion, are different. There are also some late rules which are not phonetic, if Booij and Rubach are correct about the analysis of Polish. I will not pursue any other linguistic issues in this dissertation, except to make the passing remark that the spreading and transform cases could perhaps be unified in some measure if the transforms were all seen as non-contrastive features—with  $\oplus$  now bearing the burden of combining all feature interpretations in a uniform way; the deletion analysis of the assimilation cases would perhaps remain somewhat mysterious.

I finish by reflecting on the issues with which I began. I began this dissertation by assuring linguist readers that all of their apprehensions about statistical approaches to language were misapprehensions, and that, when statistics is assigned its proper role, there is absolutely nothing subversive about its use: no one is taking away your UG.

In fact, some of the best arguments *against* “general-purpose learning” come from statistics. For one, the *bias–variance tradeoff* refers to the fact that learning can give poor results (high prediction error) for one of two reasons: too much bias (hewing so firmly to

one type of solution that the pattern is missed); or too much variance (estimated model is too sensitive to small changes in the data, and thus makes incorrect predictions on new data). More complex models will in general be able to capture more nuances of the data, at the expense of the ability to generalize. Apart from militating against frameworks that are too flexible, this has an immediate consequence when related back to Chapter 2: if our grammatical theories really do have adjustable complexity, (and I have argued that in an important way they all do anyhow), then the Chomsky’s focus on simplicity in the 1960s (and indeed Pāṇini’s much earlier—see Kiparsky 2002) is probably more important to the learner than has been previously realized; and, if they do not have adjustable complexity, then we make such a move at our peril. Exemplar-based approaches should be seen in this light; see Geman, Bienenstock & Doursat 1992 and Hastie, Tibshirani & Friedman 2009.

The *no free lunch theorem* (Wolpert 1996) states that there is no approach to learning that will do universally well: the expected classification error depends on how close the posterior of the model is to the posterior implied by the process actually generating the data. This sounds obvious, but the consequence is that if one averages over all learning problems, all algorithms perform exactly the same. The one glimmer of hope for the grammar learning problem is that the original theorem is stated only for supervised learning (access to the “correct” answers during learning); no one has, to my knowledge, correctly reformulated the theorems for the unsupervised case, but it does not seem to me that there is anything in the theorem which could not be reformulated when extended to unsupervised learning. We should not come away with the message that the field of statistics and machine learning exists simply to prove that learning is hard: the point is simply that there is well-developed formal theory already on the market with the power to answer

the kinds of big-picture questions linguists care about. Some things which appear easy turn out to be difficult when studied carefully, and some things which appear difficult turn out to be easy.

It should also, I hope, be clear by now that the proper role of statistics in linguistic theory has nothing directly to say about the other types of “numerism” sometimes perceived by linguists to be pernicious, namely, *gradience in grammar* and *frequency effects*. This is not to say that these things are not relevant, but they are not about inference per se. Gradience in grammar refers to various things. One thing it sometimes refers to is *gradient judgments*; however, as thorough treatments of the issue tend to point out, (see, for example, Keller 2000), gradient grammatical judgments—and, more to the point, gradient grammaticality—have been in the literature and have been taken to be important from the very beginnings of generative grammar (Chomsky 1975, Chomsky 1957, Chomsky & Halle 1968). Gradient judgments do not necessarily have anything to do with statistical theory—and they certainly do not need to be tied in any close way to frequency. They are important to the extent that judgments reflect “degree of grammaticality” and to the extent that “degree of grammaticality” is really the same evaluation as “goodness of fit,” (or the foundation for it), which is an important tenet of Bayesian reductionism; they need not necessarily be the same thing, however. One thing that has not been extensively considered since the days of the evaluation metrics, and which has been given new life here, is the issue of gradient *biases*—that is, a gradient UG. This is the default state of affairs for the Bayesian learner.

Gradience in grammar also sometimes refers to fine phonetic detail being available where it is not normally thought to go, namely, in lexical storage. Some studies seem to

show this (Andruski, Blumstein & Burton 1994, McMurray, Tanenhaus & Aslin 2002). However, adding phonetic detail, contrary to the intuition, does not make anything about the way language works “simpler” or “more transparent” for the learner. This merely increases the complexity of the learning problem, and raises the possibility that learners would acquire phonetic details and then fail to be able to recognize small variations in pronunciation (overfitting—see above). Nor does the existence of phonetic detail say anything about whether categorical information is stored also. If one general lesson can be taken away from the analysis of learning in this dissertation, it is that the scope of the learning problem is quite vast no matter what, and, to make it manageable, a hierarchical Bayesian learner benefits from adjustable complexity to allow additional information to be learned only when necessary. Categorical phonological patterns still do exist, although, as I pointed out in the introduction, far too few phonetic studies have explored the empirical facts of categoricity given the availability of more and more large corpora in recent years.

Finally, frequency effects are not the same as the capacity to reason statistically, and to the extent that reasoning statistically implies some kind of tracking of frequencies, this does not imply that these frequencies will show persistent effects in processing. One can easily imagine how online processing would benefit from frequency-sensitivity (better prediction, for example), but this does not necessarily lead us to any deep conclusions. In particular, arguments that frequency effects imply “whole-item storage” tend to jump the gun somewhat. Frequency effects imply that frequency of something is, in some sense, tracked, but the link to lexical storage is indirect, and the link to lexical storage of entire items is tenuous. Careful thought along these lines has already helped to disentangle the various sources of frequency effects (Pylkkänen, Stringfellow & Marantz 2002; Embick

& Marantz 2005). I would add only that, like coarse storage of phonetic categories and smaller numbers of categories, decomposed morphological representations will generally reap the general benefits of there being less to learn, and fewer ways to fail. This benefit due to reduction in the complexity of inference is not just an intuition; it can be stated mathematically as the Bayesian Occam's Razor.

In short: careful re-examination of any issue in linguistics through the lens of statistical inference, I predict, reveal effects which are crucially due to learning, and which crucially need to be tied to both gradient biases and gradient goodness-of-fit-evaluation. The degree of precision with which we can reason about such issues with statistical theory in hand makes it indispensable.

## 5.4 Main findings

I summarize again the main points and (logical) findings of this dissertation:

### 1. Statistics in linguistics

- (a) *Brief argument:* Simplicity-based evaluation of grammars is implicit in the standard practice of linguistics, and we should pay close attention to considerations of simplicity for this reason
- (b) *Main argument:* Bayesian inference can derive a preference for simpler grammars as long as grammars have some kind of structure to them; the linking principle is a Minimalist principle for prior construction called the *Optimal Measure Principle*

## 2. Modelling phonetic category learning

- (a) *Main argument:* A learning model that treats allophones as phonetic grammar outperforms a learning model that attempts to learn canonical surface representations on a simple problem using real phonetic data
- (b) *Excursus:* Statistical learning models for phonetic category learning can and should be fruitfully extended to handle learning with phonetic features

## 3. The phonetics–phonology interface

- (a) *Main idea:* Allophony is due to context-dependent phonetic transforms which take categorical phonological representations as their sole input
  - i. *Corollary:* Canonical surface representations do not exist, forcing us to reanalyze certain patterns
  - ii. *Corollary:* Allophonic processes should not feed each other
    - A. *Argument:* Certain processes involving late phonetic spreading of a feature must be qualitatively different from other phonetic rules
    - B. *Conjecture:* These rules involve delinking of phonetic features, and this is a qualitatively different process from phonetic transform application
- (b) *Argument:* Complementary distribution will encourage the learner to assign a pattern to the phonetic component of grammar
  - i. *Corollary:* Allophony is late, confirming the empirical generalizations going back to the 1970s



ii. *Corollary*: Incomplete neutralization is just allophony

(c) *Argument*: Allophony is unrecoverable

i. *Corollary*: Allophony **must** be late, explaining the empirical generalization

ii. *Corollary*: The cyclic spellout interpretation of phases or other types of cyclic level ordering is compatible with an interpretation as phonetic interpretation, so long as non-allophonic information remains recoverable

## Bibliography

- Abramson, Arthur S. & Leigh Lisker. 1970. Discriminability Along the Voicing Continuum: Cross-Language Tests. *Proceedings of Sixth International Conference of Phonetic Sciences*. 569–573.
- Adriaans, Frans & René Kager. 2010. Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language* 62. 311–331. <http://www.sciencedirect.com/science/article/pii/S0749596X09001120>.
- Albright, Adam & Bruce Hayes. 1999. An Automated Learner for Phonology and Morphology.
- Andruski, Jean, Sheila E Blumstein & Martha Burton. 1994. The effect of subphonetic differences on lexical access. *Cognition* 52(3). 163–187.
- Antoniak, Charles E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2(6). 1152–1174. <http://www.jstor.org/stable/2958336>.
- Archangeli, Diana. 1984. *Underspecification in Yawelmani Phonology and Morphology*. MIT PhD Dissertation.
- Aronoff, Mark. 1974. *Word-Structure*. MIT PhD Dissertation.

- Avery, J. Peter. 1996. *The Representation of Voicing Contrasts*. University of Toronto PhD Dissertation.
- Avery, Peter & Keren Rice. 1989. Segment Structure and Coronal Underspecification. *Phonology* 6(2). 179–200. <http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=2395536>.
- Bach, Emmon & R.T. Harms. 1972. How do languages get crazy rules? In Robert Stockwell & Ronald Macaulay (eds.), *Linguistic change and generative theory*, 1–21. Bloomington, IN: Indiana University Press.
- Bar-Hillel, Yehoshua, Chaim Gaifman & Eli Shamir. 1963. On categorial and phrase-structure grammars. *Bulletin of the Research Council of Israel* F(9). 1–16.
- Berko, Jean. 1958. The child's learning of English morphology. *Word* 14. 150–177. <http://books.google.com/books?hl=en&lr=&id=a1qJZpDU9YUC&oi=fnd&pg=PA253&dq=The+child%27s+learning+of+English+morphology&ots=NCdkgaP9f4&sig=21ExUTBP7V2KS53n8czpcdENoVs>.
- Bermúdez-Otero, Ricardo. 2013. The stem-level syndrome.
- Bermúdez-Otero, Ricardo & April McMahon. 2006. English phonology and morphology. In Bas Aarts & April McMahon (eds.), *The handbook of english linguistics*, 382–410. Oxford, UK: Blackwell.
- Berwick, Robert C, Paul Pietroski, Beracah Yankama & Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35(7). 1207–42. <http://www.ncbi.nlm.nih.gov/pubmed/21824178>.
- Berwick, Robert C. & Amy Weinberg. 1984. *The Grammatical Basis of Linguistic Performance*. Cambridge, MA: MIT Press.

- Blanchard, Daniel & Jeffrey Heinz. 2008. Improving Word Segmentation by Simultaneously Learning Phonotactics. *CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning* (August). 65–72.
- Bliese, Loren F. 1981. *A Generative Grammar of Afar*. Arlington, TX: Summer Institute of Linguistics.
- Bobaljik, Jonathan & Dianne Jonas. 1996. Subject positions and the roles of TP. *Linguistic Inquiry* 27. 195–236.
- de Boer, Bart & Patricia K. Kuhl. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online* 4(4). 129–134. <http://link.aip.org/link/ARLOFJ/v4/i4/p129/s1&Agg=doi>.
- Boersma, Paul. 2001. Empirical tests of the gradual learning algorithm. *Linguistic inquiry* 32. 45–76. <http://www.mitpressjournals.org/doi/abs/10.1162/002438901554586>.
- Boersma, Paul & Jo Pater. 2007. Constructing constraints from language data: The case of Canadian English diphthongs.
- Booij, Geert. 1977. *Dutch Morphology: A Study of Word Formation in Generative Grammar*. Dordrecht: Foris Publications.
- Booij, G & J Rubach. 1987. Postcyclic versus postlexical rules in Lexical Phonology. *Linguistic Inquiry* 18. 11–44. <http://www.jstor.org/stable/10.2307/4178523>.
- Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume & Keith Johnson. 2008. The influence of allophony vs contrast on perception: The case of Spanish and English. In Peter Avery, B. Elan Dresher & Keren Rice (eds.), *Phonological contrast*. The Hague: Mouton.

- Brame, Michael K. 1974. The Cycle in Phonology: Stress in Palestinian, Maltese, and Spanish. *Linguistic Inquiry* 5(1). 39–60.
- Braver, Aaron. 2011. Incomplete neutralization in American English flapping: A production study. *University of Pennsylvania Working Papers in Linguistics* 17(1). 1–11.
- Bresnan, Joan W. 1972. Stress and syntax: a reply. *Language* 48(2). 326–342. <http://www.jstor.org/stable/10.2307/412138>.
- Browman, Catherine P & Louis Goldstein. 1993. Dynamics and Articulatory Phonology. *Haskins Laboratories Status Report on Speech Research* SR-113. 51–62.
- Brown, P. J., M. Vannucci & T. Fearn. 1998. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(3). 627–641. <http://doi.wiley.com/10.1111/1467-9868.00144>.
- Bush, Christopher A. & Steven N. MacEachern. 1996. A semiparametric Bayesian model for randomised block designs. *Biometrika* 83(2). 275–285. <http://biomet.oupjournals.org/cgi/doi/10.1093/biomet/83.2.275>.
- Charles-Luce, Jan, Kelly Dressler & Elvira Ragonese. 1999. Effects of semantic predictability on children's preservation of a phonemic voice contrast. *Journal of Child Language* 26. 505–530.
- Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.
- Chomsky, Noam. 1964. Current Issues in Linguistic Theory. In Jerry Fodor & Jerrold Katz (eds.), *The structure of language: readings in the philosophy of language*. Englewood Cliffs, NJ: Prentice Hall. <http://en.scientificcommons.org/48023268>.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon.

- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris Publications.
- Chomsky, Noam. 1986. *Knowledge of Language*. New York: Praeger.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
- Chomsky, Noam. 2001. Derivation by phase. In *Ken Hale: a life in language*. Cambridge, MA: MIT Press.
- Chomsky, Noam & Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1. 97–214.
- Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. New York, NY: Harper & Row.
- Cinque, Guglielmo. 1993. A null theory of phrase and compound stress. *Linguistic inquiry* 24. 239–297. <http://www.jstor.org/stable/10.2307/4178812>.
- Clark, R & Ian Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24(2). 299–345. <http://www.jstor.org/stable/10.2307/4178813>.
- Cohn, Abigail. 1993. Nasalization in English: Phonology or Phonetics. *Phonology* 10(1). 43–82.
- Cohn, Abigail C. 1990. *Phonetic and phonological rules of nasalization*. UCLA PhD thesis.
- Collet, Pierre, Antonio Galves & Arturo Lopes. 1995. Maximum Likelihood and Minimum Entropy Identification of Grammars. *CoRR* cmp-lg/950.
- Compton, Richard & Christine Pittman. 2010. Word-Formation by phase in Inuit. *Lingua* 120(9). 2095–2318.

- Cornell, Sonia a, Aditi Lahiri & Carsten Eulitz. 2011. “What you encode is not necessarily what you store”: evidence for sparse feature representations from mismatch negativity. *Brain Research* 1394. 79–89. <http://www.ncbi.nlm.nih.gov/pubmed/21549357>.
- Cox, Richard T. 1946. Probability, frequency and reasonable expectation. *American Journal of Physics* 14(1). 1–13. <http://www.cco.caltech.edu/~jimbeck/summerlectures/references/ProbabilityFrequencyReasonableExpectation.pdf>.
- Crain, Stephen & Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language* 63. 522–543.
- Dawid, A.P. 1981. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* 68(1). 265. <http://biomet.oxfordjournals.org/content/68/1/265.short>.
- Dell, Gary. 1986. A spreading activation theory of retrieval in language production. *Psychological Review* 93. 283–321.
- Denis, Derek & Mark Pollard. 2008. An Acoustic Analysis of The Vowel Space of Inuktitut. *Inuktitut Linguistics Workshop*.
- Dillon, Brian, Ewan Dunbar & William Idsardi. 2013. A single-stage approach to learning phonological categories: insights from inuktitut. *Cognitive Science* 37(2). 344–77.
- Dorais, Louis-Jacques. 1986. Inuktitut surface phonology: A trans-dialectal survey. *International Journal of American Linguistics* 52(1). 20–53. <http://www.jstor.org/stable/1265501>.

- Dowe, D. L., S. Gardner & G. Oppy. 2007. Bayes not Bust! Why Simplicity is no Problem for Bayesians. *The British Journal for the Philosophy of Science* 58(4). 709–754.  
<http://bjps.oxfordjournals.org/cgi/doi/10.1093/bjps/axm033>.
- Dresher, BE. 2009a. Stress assignment in Tiberian Hebrew. In Eric Raimy & Charles Cairns (eds.), *Architecture and representations in phonological theory*. Cambridge, MA: MIT Press. <http://books.google.com/books?hl=en&lr=&id=BFofF9bxA1sC&oi=fnd&pg=PA213&dq=Stress+Assignment+in+Tiberian+Hebrew&ots=HjYCP22y8Z&sig=kMxLw2sQeawMz0aMtAm3JkCnX8A>.
- Dresher, B. Elan. 2009b. *The Contrastive Hierarchy in Phonology*. Cambridge, UK: Cambridge Univ Pr.
- Dresher, B. Elan & Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 34(2). 137–195.
- Dunbar, Ewan. 2008. *The Acquisition of Morphophonology Under a Derivational Theory: A Basic Framework and Simulation Results* (MA Thesis). University of Toronto.
- Dunbar, Ewan & William Idsardi. The Acquisition of Phonological Inventories. In Jeff Lidz, William Snyder & Joe Pater (eds.), *Oxford handbook of developmental linguistics*. Oxford: Oxford University Press.
- Dyck, Carrie. 1995. *Constraining the phonology-phonetics interface: With exemplification from Spanish and Italian dialects*. University of Toronto PhD Thesis.
- Elsner, Micha, Sharon Goldwater & Jacob Eisenstein. 2012. Bootstrapping a Unified Model of Lexical and Phonetic Acquisition. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (July). 184–193.



- Embick, David & Alec Marantz. 2005. Cognitive neuroscience and the English past tense: comments on the paper by Ullman et al. *Brain and language* 93(2). <http://www.ncbi.nlm.nih.gov/pubmed/15781308>.
- Ernestus, Mirjam & R. Harald Baayen. 2006. The functionality of incomplete neutralization in Dutch: The case of past-tense formation. In *Laboratory phonology* 8, 27–49. Berlin: Mouton De Gruyter. <http://books.google.com/books?hl=en&lr=&id=e86YANpgyisC&oi=fnd&pg=PA27&dq=The+functionality+of+incomplete+neutralization+in+Dutch:+The+case+of+past-tense+formation&ots=QQtuGE5cpC&sig=9mPEstqAP152okmMY4FupwquOMM>.
- Ernestus, Mirjam, Mybeth Lahey, Femke Verhees & R Harald Baayen. 2006. Lexical frequency and voice assimilation. *The Journal of the Acoustical Society of America* 120(2). 1040–51. <http://www.ncbi.nlm.nih.gov/pubmed/16938990>.
- Evans, Bronwen G. & Paul Iverson. 2004. Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America* 115(1). 352. <http://link.aip.org/link/JASMAN/v115/i1/p352/s1&Agg=doi>.
- Feldman, Naomi, Thomas Griffiths & James Morgan. 2009. Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. 2208–2213.
- Ferguson, Thomas S. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1(2). 209–230. <http://www.jstor.org/stable/2958008>.
- Flemming, Edward. 1995. *Auditory representations in phonology*. UCLA PhD Dissertation.

- Foraker, Stephani, Terry Regier, Naveen Khetarpal, Amy Perfors & Joshua Tenenbaum. 2009. Indirect evidence and the poverty of the stimulus: the case of anaphoric one. *Cognitive Science* 33(2). 287–300. <http://www.ncbi.nlm.nih.gov/pubmed/21585472>.
- Forster, Malcolm & Elliott Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science* 45(1). 1–35. <http://bjps.oxfordjournals.org/cgi/doi/10.1093/bjps/45.1.1><http://bjps.oxfordjournals.org/content/45/1/1.short>.
- Foulkes, Paul, James M Scobbie & Dominic Watt. 2010. Sociophonetics. *The Handbook of Phonetic Sciences, Second Edition*. 703–754.
- Fourakis, Marios & Gregory Iverson. 1984. On the 'incomplete neutralization' of German final obstruents. *Phonetica* 41. 140–149.
- Fowler, Carol A. 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14. 3–28.
- van Fraassen, Bas. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.
- Frisch, Stefan a. & Richard Wright. 2002. The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics* 30(2). 139–162. <http://linkinghub.elsevier.com/retrieve/pii/S0095447002901762>.
- Fromkin, Victoria. 1973. *Speech Errors as Linguistic Evidence*. The Hague: Mouton.
- Gagliardi, Annie. 2012. *Input and Intake in Language Acquisition*. University of Maryland, College Park PhD Dissertation.

- Gagliardi, Annie, Erin Bennett, Jeffrey Lidz & Naomi H Feldman. 2012. Children's Inferences in Generalizing Novel Nouns and Adjectives. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.
- Gagliardi, Annie & Jeffrey Lidz. In press. Statistical Insensitivity in the Acquisition of Tsez Noun Classes. *Language*.
- Geman, S, E Bienenstock & R Doursat. 1992. Neural networks and the bias variance dilemma. *Neural Computation* 4(1). 1–58. <http://www.mitpressjournals.org/doi/abs/10.1162/neco.1992.4.1.1>.
- Geman, Stuart & Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6). 721–741.
- Goldsmith, John. 1976. An overview of autosegmental phonology. *Linguistic Analysis* 2(1). 23–68.
- Goodman, N. 1955. *Fact, Fiction and Forecast*. Cambridge, MA: Harvard Univ Press.
- Goro, Takuya. 2007. *Language-Specific Constraints on Scope Interpretation in First Language Acquisition*. University of Maryland, College Park PhD Dissertation.
- Gouskova, Maria. 2003. *Deriving Economy: Syncope in Optimality Theory*. University of Massachusetts, Amherst PhD.
- Griffiths, Thomas L & Zoubin Ghahramani. 2006. Infinite Latent Feature Models and the Indian Buffet Process. *Advances in Neural Information Processing Systems* 18.
- Grijzenhout, Janet & Martin Krämer. 2000. Final devoicing and voicing assimilation in Dutch derivation and cliticization. In Barbara Stiebels & Dieter Wunderlich (eds.), *Studia grammatica 45: lexicon in focus*, 55–82. Berlin: Akademie Verlag.

- Gulian, Margarita, Paola Escudero & Paul Boersma. 2007. Supervision Hampers Distributional Learning of Vowel Contrasts. *ICPhS* (August). 1893–1896.
- Hale, Mark & Charles Reiss. 2008. *The Phonological Enterprise*. Oxford: Oxford University Press.
- Hall, Daniel Currie. 2007. *The Role and Representation of Contrast in Phonological Theory*. University of Toronto PhD.
- Halle, M & Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In Ken Hale & Samuel Jay Keyser (eds.), *The view from building 20: essays in linguistics in honor of sylvain bromberger*. Cambridge, MA: MIT Press. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Distributed+Morphology+and+the+pieces+of+inflection#0>.
- Halle, Morris. 1959. *The Sound Pattern of Russian*. The Hague: Mouton.
- Halle, Morris & Michael Kenstowicz. 1991. The Free Element Condition and Cyclic versus Noncyclic Stress. *Linguistic Inquiry* 22(3). 457–501.
- Halle, Morris & K. P. Mohanan. 1985. Segmental phonology of Modern English. *Linguistic inquiry* 16(1). 57–116. <http://www.jstor.org/stable/10.2307/4178420>.
- Hall, Kathleen Currie & E Allyn Smith. 2006. Finding vowels without phonology? *Montreal-Ottawa-Toronto Phonology Workshop* (February).
- Hamann, Silke. 2003. *The Phonetics and Phonology of Retroflexes*. Utrecht: LOT Press.
- Hastie, Trevor, Rob Tibshirani & Jerome Friedman. 2009. *The Elements of Statistical Learning*. New York: Springer.

- Hayes, Bruce. 1984. The phonetics and phonology of Russian voicing assimilation. In Mark Aronoff & Richard Oehrle (eds.), *Language sound structure*, 318–328. Cambridge, MA: MIT Press.
- Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.
- Heinz, Jeffrey. 2013. Computational theories of learning and developmental psycholinguistics. In Jeffrey Lidz, William Snyder & Joe Pater (eds.), *The cambridge handbook of developmental linguistics*. Cambridge: Cambridge University Press.
- Heinz, Jeffrey & William Idsardi. 2011. Sentence and word complexity. *Science* 333(6040). 295–297. <http://www.ncbi.nlm.nih.gov/pubmed/21764736>.
- Henderson, Leah, Noah D Goodman, Joshua B Tenenbaum & James F Woodward. 2010. The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. 77(2). 172–200.
- Herd, Wendy, Allard Jongman & Joan Sereno. 2010. An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. *Journal of Phonetics* 38(4). 504–516. <http://linkinghub.elsevier.com/retrieve/pii/S0095447010000458>.
- Hillenbrand, J., L.A. Getty, M.J. Clark & K. Wheeler. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97(5). 3099–3111. <http://winnie.kuis.kyoto-u.ac.jp/members/okuno/Lecture/05/Hearing/Hillenbrand-JASA-97-5-3099-3111.pdf>.
- Hjort, NL. 1990. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* 18. 1259–1294. <http://www.jstor.org/stable/10.2307/2242052>.

- Hoijer, H. 1949. Tonkawa: An Indian language of Texas. In C Osgood (ed.), *Linguistic structures of native america*. New York: Viking Fund Publications in Anthropology 6.
- Hooper, Joan. 1976. *An Introduction to Natural Generative Phonology*. New York: Academic Press.
- Idsardi, William. 2006. Canadian Raising, Opacity, and Rephonemicization. *The Canadian Journal of Linguistics / La revue canadienne de linguistique* 51(2). 119–126.  
[http://muse.jhu.edu/content/crossref/journals/canadian\\_journal\\_of\\_linguistics/v051/51.2idsardi.pdf](http://muse.jhu.edu/content/crossref/journals/canadian_journal_of_linguistics/v051/51.2idsardi.pdf).
- Jackendoff, Ray. 1977. *X-Bar Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press.
- Jakobson, Roman. 1929. Remarques sur l'évolution phonologique du russe comparée à celle des autres langues slaves. In *Selected works*.
- Jakobson, Roman. 1941. *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala: Almqvist & Wiksells.
- Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. *Preliminaries to speech analysis: The distinctive features*. Cambridge, MA: MIT Press.
- Jakobson, Roman & Morris Halle. 1956. *Fundamentals of Language*. The Hague: Mouton.
- Jarosz, Gaja. 2006. *Rich Lexicons and Restrictive Grammars: Maximum Likelihood Learning in Optimality Theory*. Johns Hopkins University PhD thesis.
- Jarosz, Gaja. 2011. The Roles of Phonotactics and Frequency in the Learning of Alternations. *BUCLD 35 Proceedings*.
- Jaynes, E. T. 2003. *Probability Theory: The Logic Of Science*. Cambridge: Cambridge University Press.

- Jessen, Michael. 1998. *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam: John Benjamins.
- Johnson, C. Douglas. 1972. *Formal aspects of phonological description*. The Hague: Mouton.
- Jones, Matt & Bradley C Love. 2011. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and brain sciences* 34(4). <http://www.ncbi.nlm.nih.gov/pubmed/21864419>.
- Kaplan, Ronald M & Martin Kay. 1994. Regular Models of Phonological Rule Systems. *Computational Linguistics*.
- Kaye, Jonathan. 1990. What ever happened to Dialect B? In Joan Mascaró & Marina Nespor (eds.), *Grammar in progress: glow essays for henk van riemsdijk*, 259–263. Dordrecht: Foris Publications.
- Kazanina, Nina, Colin Phillips & William Idsardi. 2006. The influence of meaning on the perception of speech sounds. *The Journal of the Acoustical Society of America* 103(30). 11381–11386. <http://www.ncbi.nlm.nih.gov/pubmed/16849423>.
- Kean, Mary-Louise. 1975. *The theory of markedness in generative grammar*. MIT PhD thesis. <http://en.scientificcommons.org/30563611>.
- Keating, Patricia A. 1988. Underspecification in phonetics. *Phonology* 3(2). 275–292.
- Keller, Frank. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. University of Edinburgh PhD Dissertation.

- Kemp, Alan. 1994. Phonetic transcription: History. In R. E. Asher & E. J. A. Henderson (eds.), *The encyclopedia of language and linguistics: volume 6*, 3040–3051. Oxford, UK: Pergamon Press.
- Keyser, Samuel Jay & Kenneth N. Stevens. 2006. Enhancement and Overlap in the Speech Chain. *Language* 82(1). 33–63. <http://muse.jhu.edu/content/crossref/journals/language/v082/82.1keyser.pdf>.
- Kim, Hyunsoon & Allard Jongman. 1996. Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean. *Journal of Phonetics* 24. 295–312.
- Kiparsky, Paul. 1971. Historical linguistics. In W.O. Dingwall (ed.), *A survey of linguistic science*. College Park, MD: University of Maryland Linguistics Program.
- Kiparsky, Paul. 1982. From Cyclic Phonology to Lexical Phonology. In Harry van der Hulst & Norval Smith (eds.), *The structure of phonological representations*, 131–175. Dordrecht: Foris Publications.
- Kiparsky, Paul. 1985. Some consequences of Lexical Phonology. *Phonology* 2. 85–138.
- Kiparsky, Paul. 1991. Economy and the Construction of the Śivasūtras. In M. M. Deshpande & S. Bhate (eds.), *Paninian studies*. Ann Arbor, MI.
- Kiparsky, Paul. 2000. Opacity and Cyclicity. *The Linguistic Review* 17. 351–367.
- Kiparsky, Paul. 2002. On the Architecture of Pāṇini's Grammar.
- Kisseberth, Charles. 1970. Vowel elision in Tonkawa and derivational constraints. In J. Sadock & A. Vanek (eds.), *Studies presented to Robert B. Lees by his students*. Champaign, IL: Linguistic Research.
- Knaus, Johannes, Richard Wiese & Ulrike Domahs. 2011. Secondary stress is distributed rhythmically within words: an EEG study on German. *ICPhS* (August). 1114–1117.



- Kuhl, Patricia K. 1991. Human adults and human infants show a “perceptual magnet” effect for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50(2). 93–107.
- Ladefoged, Peter. 2005. Features and parameters for different purposes. *UCLA Working Papers in Phonetics* 104. 1–13. [http://www.linguistics.ucla.edu/faciliti/workpapph/104/1-PL\\_lsa\\_january\\_2005.pdf](http://www.linguistics.ucla.edu/faciliti/workpapph/104/1-PL_lsa_january_2005.pdf).
- Lahiri, Aditi & B. Elan Dresher. 1984. Diachronic and synchronic implications of declension shifts. *The Linguistic Review* 3. 141–163.
- Lahiri, Aditi & Henning Reetz. 2002. Underspecified Recognition. *Labphon* 7.
- Lahiri, Aditi & Henning Reetz. 2010. Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics* 38(1). 44–59. <http://linkinghub.elsevier.com/retrieve/pii/S0095447010000033>.
- Lamb, Sydney. 1964. On alternation, transformation, realization, and stratification. *Monograph Series on Languages and Linguistics* 17. 105–22.
- Lasnik, Howard & Juan Uriagereka. 2002. On the poverty of the challenge. *The Linguistic Review* 19. 147–150. [http://www.degruyter.com/dg/viewarticle.fullcontentlink:pdfeventlink/contentUri?format=INT&t:ac=j\\$002ftlir.2002.19.issue-1-2\\$002ftlir.19.1-2.147\\$002ftlir.19.1-2.147.xml](http://www.degruyter.com/dg/viewarticle.fullcontentlink:pdfeventlink/contentUri?format=INT&t:ac=j$002ftlir.2002.19.issue-1-2$002ftlir.19.1-2.147$002ftlir.19.1-2.147.xml).
- Leben, William. 1973. *Suprasegmental Phonology*. MIT PhD Dissertation.
- Legate, Julie & Charles Yang. 2002. Empirical re-assessment of stimulus poverty. *The Linguistic Review* 19. 151–162.

- Liberman, Alvin M. & Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition* 21(1). 1–36. <http://www.ncbi.nlm.nih.gov/pubmed/4075760>.
- Liberman, Mark & Janet Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In Mark Aronoff & Richard Oehrle (eds.), *Language sound structure*, 157–233. Cambridge, MA: MIT Press. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Intonational+invariance+under+changes+in+pitch+range+and+length#0>.
- Liberman, Mark & A. Prince. 1977. On Stress and Linguistic Rhythm. *Linguistic inquiry* 8(2). 249–336. <http://www.jstor.org/stable/10.2307/4177987>.
- MacEachern, Steven N. 1994. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* 23(3). 727–741.
- MacEachern, Steven N. 1999. Dependent Dirichlet processes. [http://stat.columbia.edu/\\$\sim\\$porbanz/talks/MacEachern2000.pdf](http://stat.columbia.edu/$\sim$porbanz/talks/MacEachern2000.pdf).
- MacKay, David. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.
- MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt & Co.
- Martin, Joel, Howard Johnson, Benoit Farley & Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. *Proceedings of the HLT-NAACL 2003 Work-*

- shop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond* (June). 115–118. <http://portal.acm.org/citation.cfm?doid=1118905.1118925>.
- Mascaró, Joan. 1976. *Catalan Phonology and the Phonological Cycle*. MIT PhD thesis.
- Maye, Jessica, Janet F. Werker & LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82(3). B101–B111. <http://www.ncbi.nlm.nih.gov/pubmed/11747867>.
- McCarthy, John. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17. 207–263.
- McCarthy, John J. 1999. Sympathy and phonological opacity. *Phonology* 16. 331–399. [http://journals.cambridge.org/abstract\\_S0952675799003784](http://journals.cambridge.org/abstract_S0952675799003784).
- McMurray, Bob, Richard N Aslin & Joseph C Toscano. 2009. Statistical learning of phonetic categories: insights from a computational approach. *Developmental science* 12(3). 369–78. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2742678&tool=pmcentrez&rendertype=abstract>.
- McMurray, Bob, Michael K Tanenhaus & Richard N Aslin. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86(2). B33–42. <http://www.ncbi.nlm.nih.gov/pubmed/12435537>.
- Mielke, Jeff. 2008. *The Emergence of Distinctive Features*. Oxford, UK: Oxford Univ Press.
- Mielke, Jeff, Mike Armstrong & Elizabeth Hume. 2003. Looking Through Opacity. *Theoretical Linguistics* 29. 123–139.

- Miller, Kurt T, Thomas L Griffiths & Michael I Jordan. 2009. The Phylogenetic Indian Buffet Process : A Non-Exchangeable Nonparametric Prior for Latent Features. *Advances in Neural Information Processing Systems* 22.
- Morén, Bruce. 2007. Phonological Segment Inventories and their Phonetic Variation: a substance-free approach. *Generative Linguistics in the Old World (GLOWXXX)*.
- Niyogi, P & R C Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61(1-2). 161–93. <http://www.ncbi.nlm.nih.gov/pubmed/8990971>.
- Noske, Roland. 1993. *A Theory of Syllabification and Segmental Alternation: With Studies on the Phonology of French, German, Tonkawa and Yawelmani*. Tübingen: Max Niemeyer.
- van Oostendorp, Marc. 2006. Incomplete devoicing in formal phonology.
- Padgett, Jaye. 2003. Contrast and post-velar fronting in Russian. *Natural Language and Linguistic Theory* 21(1). 39–87. <http://link.springer.com/article/10.1023/A%3A1021879906505>.
- Paradis, Carole. 1988. On Constraints and Repair Strategies. *The Linguistic Review* 6. 71–97.
- Pearl, Lisa. 2007. *Necessary Bias in Natural Language Learning*. University of Maryland, College Park PhD Dissertation.
- Pearl, Lisa, Sharon Goldwater & Mark Steyvers. 2010. How ideal are we? Incorporating human limitations into Bayesian models of word segmentation. *Proceedings of the 34th Annual Boston University Conference on Child Language Development*. Cascadia Press, Somerville. <http://homepages.inf.ed.ac.uk/sgwater/papers/bucl09-onlineseg.pdf>.

- Pearl, Lisa & Jeffrey Lidz. 2009. When domain-general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development* 5. 235–265. <http://www.tandfonline.com/doi/abs/10.1080/15475440902979907>.
- Pentus, Mati. 1993. Lambek grammars are context-free. *Logic in Computer Science*. 429–433.
- Peperkamp, Sharon. 2003. Phonological acquisition: recent attainments and new challenges. *Language and speech* 46(Pt 2-3). 87–113. <http://www.ncbi.nlm.nih.gov/pubmed/14748441>.
- Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal & Emmanuel Dupoux. 2006. The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* 101(3). B31–841. <http://www.ncbi.nlm.nih.gov/pubmed/16364279>.
- Peperkamp, Sharon, Michèle Pettinato & Emmanuel Dupoux. 2002. Allophonic Variation and the Acquisition of Phoneme Categories. *Proceedings of the 27th Annual Boston University Conference on Language Development*.
- Perfors, Amy, Joshua B Tenenbaum & Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118. 306–338. <http://www.ncbi.nlm.nih.gov/pubmed/21186021>.
- Phelps, Elaine. 1975. Iteration and disjunctive domains in phonology. *Linguistic Analysis* 1. 137–172.
- Phillips, Lawrence & Lisa Pearl. 2012. “Less is More” in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. *Proceedings of the 34th Annual Conference of the Cognitive Science Society* 2011. 863–868.

- Piggott, Glyne & Heather Newell. 2006. Syllabification, stress and derivation by phase in Ojibwa. *McGill Working Papers in Linguistics*. 1–47. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Syllabification,+stress+and+derivation+by+phase+in+Ojibwa#0>.
- Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28. 73–193.
- Pisoni, David B. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Attention, Perception, and Psychophysics* 13. 253–260. <http://www.springerlink.com/index/88V549745842656J.pdf>.
- Pitt, Mark a., Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1). 89–95. <http://linkinghub.elsevier.com/retrieve/pii/S0167639304000974>.
- Poliquin, Gabriel. 2006. *Canadian French Vowel Harmony*. Harvard PhD Dissertation. <http://roa.rutgers.edu/files/861-0906/861-POLIQVIN-0-0.PDF>.
- Port, Robert & Adam Leary. 2005. Against formal phonology. *Language* 81(4). 927–964. <http://muse.jhu.edu/journals/lan/summary/v081/81.4port.html>.
- Port, Robert & Michael L. O'Dell. 1986. Neutralization of syllable-final voicing in German. *Journal of Phonetics* 13. 455–471.
- Prince, Alan. 1975. *The Phonology and Morphology of Tiberian Hebrew*. MIT PhD Dissertation.
- Prince, Alan & Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell.

- Pye, Shizuka. 1986. Word-final devoicing in Russian. *Cambridge Papers in Phonetics and Experimental Linguistics* 5. 1–10.
- Pylkkänen, Liina, Andrew Stringfellow & Alec Marantz. 2002. Neuromagnetic Evidence for the Timing of Lexical Activation: An MEG Component Sensitive to Phonotactic Probability but Not to Neighborhood Density. *Brain and Language* 81(1-3). 666–678.  
<http://linkinghub.elsevier.com/retrieve/pii/S00939334X01925556>.
- Quine, Willard Van Orman. 1960. *Word and Object*. Cambridge, MA: MIT Press.
- Rice, KD. 1993. A reexamination of the feature [sonorant]: The status of 'sonorant obstruents'. *Language* 69(2). 308–344. <http://www.jstor.org/stable/10.2307/416536>.
- Ringen, Catherine O & Robert M Vago. 1998. Hungarian vowel harmony in Optimality Theory. *Phonology* 15. 393–416.
- Rubach, Jerzy. 1984. *Cyclic and Lexical Phonology: The Structure of Polish*. Dordrecht: Foris Publications.
- Rubach, Jerzy & Geert Booij. 1990. Syllable structure assignment in Polish. *Phonology* 7(1). 121–158. [http://journals.cambridge.org/abstract\\_S0952675700001135](http://journals.cambridge.org/abstract_S0952675700001135).
- Rumelhart, David E & James L McClelland. 1986. On learning the past tenses of English verbs. In David E Rumelhart, James L McClelland & The PDP Research Group (eds.), *Parallel distributed processing: explorations in the microstructure of cognition. volume 2: psychological and biological models*. Cambridge, MA: Bradford Books/MIT Press.

- Salor, Ö, B.L. Pellom, T. Ciloglu & M Demirekler. 2007. Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. *Computer Speech and Language* 21. 583–593.
- Sampson, Geoffrey. 2002. Exploring the richness of the stimulus. *The Linguistic Review* 19. 73–104.
- Scharinger, Mathias, Aditi Lahiri & Carsten Eulitz. 2010. Mismatch negativity effects of alternating vowels in morphologically complex word forms. *Journal of Neurolinguistics* 23(4). 383–399. <http://linkinghub.elsevier.com/retrieve/pii/S091160441000028X>.
- Scharinger, M. & A. Lahiri. 2010. Height Differences in English Dialects: Consequences for Processing and Representation. *Language and Speech* 53(2). 245–272. <http://las.sagepub.com/cgi/doi/10.1177/0023830909357154>.
- Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8. 333–343.
- Slavin, Tanya. 2012. Truncation and Morphosyntactic Structure in Ojicree. *McGill Working Papers in Linguistics* 22. 1–12. <https://secureweb.mcgill.ca/mcgwpl/sites/mcgill.ca.mcgwpl/files/slavin2012.pdf>.
- Sledd, James H. 1966. Breaking, Umlaut, and the Southern Drawl. *Language* 42(1). 18–41.
- Slowiaczek, L.M. & D.A. Dinnsen. 1985. On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics* 13(3). 325–341.



- Steriade, Donc. 1987. Redundant values. *CLS 23: Papers from the 23rd Annual Regional Meeting of the Chicago Linguistic Society. Part Two: Parasession on Autosegmental and Metrical Phonology*. 339–62.
- Stevens, Jon. 2011. Learning Object Names in Real Time with Little Data. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. 903–908.
- Stevens, Jon, John Trueswell, Charles Yang & Lila Gleitman. 2013. The Pursuit of Word Meanings. [http://www.ircs.upenn.edu/\\$\sim\\$truesweb/trueswell\\_pdfs/Stevens\\_et\\_al\\_submitted.pdf](http://www.ircs.upenn.edu/$\sim$truesweb/trueswell_pdfs/Stevens_et_al_submitted.pdf).
- Stevens, Kenneth N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111(4). 1872. <http://link.aip.org/link/JASMAN/v111/i4/p1872/s1&Agg=doi>.
- Teh, Yee Whye & Dilan Görür. 2009. Indian Buffet Processes with Power-law Behavior. *Advances in Neural Information Processing Systems* 22.
- Tenenbaum, Joshua. 1999. *A Bayesian Framework for Concept Learning*. Cambridge, MA: MIT PhD thesis.
- Thibaux, Romain & Michael I Jordan. 2007. Hierarchical Beta Processes and the Indian Buffet Process. *International Conference on Artificial Intelligence and Statistics*. 564–571.
- Uriagereka, Juan. 1999. Multiple spell-out. In Samuel D. Epstein & Norbert R. Hornstein (eds.), *Working minimalism*. Cambridge, MA: MIT Press.
- Vallabha, Gautam K, James L McClelland, Ferran Pons, Janet F Werker & Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America* 104(33).

- 13273–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1934922&tool=pmcentrez&rendertype=abstract>.
- Viau, Joshua & Jeffrey Lidz. 2011. Selective learning in the acquisition of Kannada ditransitives. *Language* 87(4). 679–714.
- Wan, I-ping & Jeri Jaeger. 1998. Speech errors and the representation of tone in Mandarin Chinese. *Phonology* 15(3). 417–461. <http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=40714>.
- Wheeler, Max. 2005. *The Phonology of Catalan*. Oxford, UK: Oxford University Press.
- Williamson, Sinead, Peter Orbanz & Zoubin Ghahramani. 2010. Dependent Indian Buffet Processes. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* 6.
- Wilson, Stephen M. 2003. A phonetic study of voiced, voiceless and alternating stops in Turkish. *CRL Newsletter* 15(1). 3–13. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+phonetic+study+of+voiced,+voiceless+and+alternating+stops+in+Turkish#0>.
- Wolpert, David. 1996. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* 8(7). 1341–1390.
- Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.
- Yip, Kenneth & Gerald Jay Sussman. 1997. Sparse Representations for Fast, One-shot learning. *Proceedings of the National Conference on Artificial Intelligence*.
- Zimmermann, Malte. 2002. *Boys Buying Two Sausages Each: On the Syntax and Semantics of Distance-Distributivity*. University of Amsterdam PhD thesis.

Zonneveld, Wim. 1983. Lexical and phonological properties of Dutch voicing assimilation. In Marcel van den Broecke, Vincent van Heuven & Wim Zonneveld (eds.), *Sound structures: studies for antonie cohen*, 297–312. Dordrecht: Foris Publications.